Katholieke Universiteit Leuven
FACULTEIT LETTEREN
Subfaculteit Taalkunde

La polysémie du vocabulaire technique

Une étude quantitative

Proefschrift ingediend tot
het behalen van de graad van
Doctor in de Taal- en Letterkunde:
Romaanse Talen

Ann Bertels

Promotoren:
Prof. dr. Béatrice Lamiroy
Prof. dr. Dirk Geeraerts

LEUVEN 2006

Dankwoord

Dit proefschrift is het resultaat van vele jaren van studie en wetenschappelijk onderzoek. Heel wat mensen hebben ertoe bijgedragen om dit werk tot een goed einde te brengen en ik zou hen langs deze weg graag willen bedanken.

Eerst en vooral bedank ik mijn promotoren, de professoren Béatrice Lamiroy en Dirk Geeraerts, voor de uitvoerige en constructieve besprekingen, voor hun kritische ingesteldheid en hun waardevolle raadgevingen. Ik bedank ook professor Dirk Speelman voor zijn inhoudelijke en statistische raadgevingen en zijn technische ondersteuning bij het programmeren in Python en bij de statistiek in R.

Mijn dank gaat ook uit naar de professoren Bernard Victorri en Cédrick Fairon, die aanvaard hebben om in de jury te zetelen. Verder bedank ik de professoren Jean Véronis, Jean-Pierre Kruth et An Carbonez voor hun nuttige raadgevingen.

Tevens bedank ik de professoren Jean Binon en Serge Verlinde, mijn collega's op het ILT, met wie ik mijn eerste stappen zette in het wetenschappelijk onderzoek (de DAFA) en die mij altijd hebben gesteund en aangemoedigd. Dank voor hun aandachtige lectuur van het manuscript en voor hun nuttige suggesties. Ik dank mijn collega Anne-marie voor de interessante besprekingen en voor het nalezen. Ook ben ik mijn andere ILT-collega's dankbaar, zeker mijn collega's van de taalgroep Frans, Stéphane, Nathalie, Dominique, Kristin, An, Bénédicte, Aurélie, Katrijn, Hilde en Evelyn, voor hun morele steun.

De leden van de onderzoeksgroep QLVL hebben mij vaak ondersteuning gegeven voor presentaties of congressen. Ik bedank daarbij Gert, José, Kris, Stefania, Sofie, Koen, Yves en Dylan voor de verrijkende discussies. In het bijzonder dank ik Stef voor het nalezen van mijn Engelse samenvatting.

Ik bedank mijn vrienden, de familie van Jan en mijn familie voor hun eindeloze aanmoedigingen. Tenslotte bedank ik Jan, omdat hij zowel de moeilijke als de vreugdevolle momenten met mij deelde en omdat hij mij aanzette om door te bijten, dag na dag.

Remerciements

Cette thèse de doctorat constitue l'aboutissement de plusieurs années de recherches et d'expérimentations. De nombreuses personnes m'ont aidée à mener à bien ce travail et je tiens à les remercier.

Tout d'abord, je remercie mes directeurs de thèse, les professeurs Béatrice Lamiroy et Dirk Geeraerts, pour les longues discussions constructives, pour leur esprit critique et leurs précieux conseils. Je remercie le professeur Dirk Speelman pour ses conseils techniques et statistiques et pour m'avoir initiée au monde passionnant de la programmation en Python et aux statistiques dans le logiciel R.

Je tiens à remercier les professeurs Bernard Victorri et Cédrick Fairon, qui ont accepté de faire partie du jury. Je remercie également les professeurs Jean Véronis, Jean-Pierre Kruth et An Carbonez pour leurs conseils prodigués à l'occasion de conférences et lors de rencontres informelles.

Je suis très reconnaissante envers les professeurs Jean Binon et Serge Verlinde, mes collègues de l'ILT, avec qui j'ai débuté mes premières recherches scientifiques (le DAFA) et qui m'ont toujours encouragée. Merci pour leurs relectures attentives et pour leurs précieuses suggestions. Un grand merci à ma collègue Anne-marie pour les discussions intéressantes et pour sa relecture. Je remercie aussi tous mes autres collègues de l'ILT et particulièrement mes collègues de français, Stéphane, Nathalie, Dominique, Kristin, An, Bénédicte, Aurélie, Katrijn, Hilde et Evelyn, pour leur soutien moral.

Les membres de l'équipe de recherche QLVL m'ont souvent soutenue pour des présentations ou des conférences. Je remercie Gert, José, Kris, Stefania, Sofie, Koen, Yves et Dylan pour les discussions enrichissantes. Je remercie Stef plus particulièrement pour sa relecture du résumé anglais.

Je tiens à remercier mes amis, la famille de Jan et ma famille, pour tous leurs encouragements. En dernier lieu, je remercie Jan, qui a partagé les moments difficiles et les moments de joie et qui m'a aidée à progresser jour après jour.

Table des matières

Liste des figures

Liste des tableaux

Introduction

La polysémie est un phénomène omniprésent dans la langue. Un calcul statistique approximatif révèle que plus de 40% des mots du Petit Robert sont polysémiques : leur entrée dans le dictionnaire comporte au moins deux subdivisions (Victorri & Fuchs 1996). En effet, la polysémie est généralement définie en termes de « pluralité de sens apparentés », correspondant à une seule unité linguistique, tant lexicale que grammaticale. Nous disposons d'un nombre limité de mots ou d'unités linguistiques pour exprimer un nombre illimité d'idées ou de notions. La plupart des études sémantiques étudient la polysémie dans la langue générale, et plus particulièrement la polysémie des unités lexicales. Rares sont les travaux consacrés à l'étude de la polysémie dans la langue spécialisée. Cela s'explique bien entendu par les efforts de normalisation de la terminologie traditionnelle, qui préconise l'idéal de monosémie et d'univocité dans la langue spécialisée des sciences et des techniques.

Récemment, on a assisté à la remise en question de l'idéal de monosémie par les partisans de la terminologie descriptive et linguistique. On a assisté en même temps à l'émergence de vastes corpus spécialisés, qui ont permis des études sémantiques à partir du contexte linguistique et qui ont abouti à l'observation de cas de polysémie dans la langue spécialisée. Ces récentes remises en question et certaines études sémantiques ponctuelles sur des corpus spécialisés nous ont incitée à étudier le phénomène de la polysémie dans la langue spécialisée à plus grande échelle. Nous procéderons dès lors à l'étude sémantique du vocabulaire technique d'un domaine spécialisé, en l'occurrence le domaine restreint des machines-outils pour l'usinage des métaux.

Remettant en question l'idéal de monosémie de la terminologie traditionnelle, nous nous demanderons si les unités lexicales spécifiques ou représentatives dans notre corpus technique sont effectivement monosémiques, tel que le préconise l'approche traditionnelle. Etant donné que nous envisageons une étude sémantique de toutes les unités lexicales spécifiques d'un corpus technique, l'automatisation et la quantification s'imposent. En effet, il est impossible d'analyser manuellement tous les contextes d'usage de toutes les occurrences de plusieurs milliers d'unités lexicales. Nous procéderons donc à une étude sémantique automatisée et nous

accorderons une valeur numérique à chaque unité lexicale analysée, en fonction de son « degré » de monosémie.

Notre étude se compose de trois grandes parties. La première partie présentera la problématique et elle comprendra deux chapitres, à savoir l'état de la question et le cadre théorique, décrits dans le premier chapitre, ainsi que les questions et les hypothèses de recherche, présentées dans le deuxième chapitre. Ensuite, la deuxième partie de notre étude constituera la partie méthodologique. Elle expliquera la constitution des corpus (chapitre 3) et les deux axes méthodologiques (chapitres 4 et 5), qui feront l'objet de plusieurs expérimentations et mises au point (chapitre 6). Finalement, la troisième partie présentera les résultats de notre étude et les interprétations linguistiques qui en découlent (chapitres 7 et 8). Elle se terminera par les conclusions et les perspectives (chapitre 9). Nous ferons régulièrement des renvois aux annexes, que nous avons préféré joindre sur support électronique. Ce CD-ROM comprend également des documents électroniques supplémentaires, notamment des listes et des visualisations plus détaillées.

Dans le *premier chapitre*, nous présenterons l'état de la question de la présente étude. Nous passerons en revue les études récentes et les travaux pertinents dans le domaine de la langue spécialisée et dans le domaine de l'analyse sémantique. A deux reprises, nous commencerons par l'explication de la dichotomie traditionnelle, à savoir la dichotomie entre mot et terme et la dichotomie entre polysémie et monosémie. Ces deux dichotomies ne s'avèrent pas toujours opérationnelles et elles sont remises en question pour plusieurs raisons. Nous tenterons d'apporter une solution alternative en adoptant une approche scalaire, c'est-à-dire un double continuum. L'état de la question permettra ensuite de situer et de justifier notre hypothèse de recherche, qui sera précisée dans le *deuxième chapitre*. Nous nous demandons si et à quel point les unités lexicales les plus spécifiques et les plus représentatives du corpus technique sont monosémiques ou polysémiques. Comme nous adoptons une double approche, à la fois quantitative et scalaire, la question principale sera celle de savoir s'il existe une corrélation entre, d'une part, le continuum de spécificité et, de l'autre, le continuum de monosémie. Nous avancerons l'hypothèse que les unités (les plus) spécifiques du corpus technique ne sont pas nécessairement (les plus) monosémiques et qu'il n'y a donc pas de corrélation positive entre le continuum de spécificité et le continuum de monosémie.

La constitution du corpus technique et du corpus de référence fera l'objet du *troisième chapitre*. Afin d'aboutir au double continuum, nous procéderons à une double analyse quantitative. Le premier axe méthodologique, expliqué dans le *quatrième chapitre*, permettra la quantification au niveau des unités lexicales et remplacera la dichotomie traditionnelle entre mot et terme. On tentera de déterminer à quel point les unités lexicales sont spécifiques ou représentatives du corpus

technique en le comparant à un corpus de référence de langue générale. Ainsi, les unités lexicales du corpus technique seront classées en fonction de leur « degré » de spécificité. Le deuxième axe méthodologique, présenté dans le *cinquième chapitre*, conduira à la quantification de la monosémie. Signalons d'ores et déjà que l'on ne peut pas « hypostasier » la monosémie. La monosémie d'une unité lexicale n'est pas une réalité objective : elle est observable à travers les occurrences de cette unité lexicale, que le contexte permet d'interpréter. Il en va de même pour la polysémie ou pour le vague. Les critères qui permettent traditionnellement de distinguer entre polysémie et vague ne sont pas toujours fiables ni convergents. En plus, ces critères ne se prêtent pas à une application opérationnelle et objective à grande échelle. Nous proposerons dès lors une analyse sémantique alternative, qui sera « quantitative ». A cet effet, le caractère monosémique d'une unité lexicale sera considéré en termes d'homogénéité sémantique. En effet, une unité lexicale monosémique se caractérise par des cooccurrences ou par des contextes sémantiquement plutôt homogènes, tandis qu'une unité lexicale polysémique apparaîtra dans des contextes sémantiquement plus hétérogènes. A partir d'une analyse des cooccurrences, et plus particulièrement en élaborant une mesure pour calculer le « degré » d'homogénéité sémantique, nous tenterons de quantifier et d'objectiver l'analyse sémantique. Evidemment, notre mesure fera l'objet de nombreuses expérimentations et mises au point méthodologiques, décrites dans le *sixième chapitre*.

Finalement, les données quantitatives de spécificité et de monosémie feront l'objet d'analyses statistiques de régression, pour toutes les unités lexicales spécifiques de notre corpus technique et pour quelques sous-ensembles. Les analyses statistiques de régression permettront d'apporter une réponse objective à la question de savoir si les unités lexicales les plus spécifiques du corpus technique sont effectivement les plus monosémiques. Dans le *septième chapitre*, nous présenterons les résultats des analyses statistiques de base, pour toutes les unités lexicales spécifiques du corpus technique. Nous formulerons également des interprétations linguistiques à partir des résultats statistiques. Dans le *huitième chapitre*, nous tenterons d'approfondir les interprétations linguistiques, en analysant différents sous-ensembles et différents sous-corpus. Le *dernier chapitre* des conclusions générales reprendra les lignes de force des chapitres précédents. Nous terminerons notre étude par quelques perspectives de recherche intéressantes pour des études sémantiques quantitatives ultérieures.

PARTIE I

Problématique

Chapitre 1
Etat de la question et cadre théorique

Dans ce premier chapitre, nous nous proposons de situer le cadre théorique de notre étude, qui se situe dans le domaine de la sémantique quantitative. Notre démarche, qui relève de la lexicologie quantitative, s'inscrit clairement dans le cadre de la terminologie descriptive, privilégiant l'exploitation de corpus de textes spécialisés pour l'étude sémantique. Toutefois, il est intéressant de s'interroger sur les différentes approches et théories terminologiques qui ont marqué l'histoire de l'étude sémantique des unités lexicales spécialisées.

Ce chapitre comprend trois parties : la première partie discutera la langue spécialisée, la deuxième sera consacrée à la sémantique et à la polysémie. Dans la troisième, nous présenterons certaines restrictions, avant de passer à l'explication des questions et hypothèses de recherche qui seront formulées dans le deuxième chapitre. Les deux premières parties de ce premier chapitre s'articuleront de la même façon : elles commenceront par présenter la dichotomie (traditionnelle), qui sera ensuite remise en question pour de nombreuses raisons. Les deux parties se terminent par la solution alternative d'une approche scalaire graduelle, que nous élaborerons dans les chapitres suivants. Les deux parties principales de ce premier chapitre, à savoir la langue spécialisée (1.1) et la sémantique (1.2), correspondent aux deux axes méthodologiques qui feront l'objet de cette étude sémantique quantitative de la langue spécialisée (Cf. chapitres 4 et 5).

1.1 LANGUE SPÉCIALISÉE

Comme cette étude vise à analyser la langue spécialisée, et plus particulièrement le vocabulaire d'un corpus technique, il convient de s'interroger sur la notion de langue spécialisée et sur les approches théoriques des dernières décennies. En effet, l'évolution de la terminologie traditionnelle à la terminologie descriptive mettra en évidence que la dichotomie entre langue générale et langue spécialisée, ou entre mot et terme, n'est pas la méthodologie idéale pour l'analyse de corpus de textes techniques.

La première section de cette partie (1.1.1) décrira la dichotomie entre langue générale et langue spécialisée. A cet effet, la langue spécialisée ou de spécialité (1.1.1.1) sera située dans le contexte de la terminologie traditionnelle (1.1.1.2), qui oppose la langue spécialisée à la langue générale, tout comme elle oppose le terme au mot (1.1.1.3). La deuxième section (1.1.2) sera consacrée à la remise en question de la dichotomie entre langue générale et langue spécialisée et décrira l'approche de la terminologie descriptive et linguistique (1.1.2.1), les différents types d'unités lexicales de la langue spécialisée (1.1.2.2), ainsi que les interactions entre langue générale et langue spécialisée (1.1.2.3). Dans la dernière section (1.1.3), nous présenterons brièvement la solution alternative, c'est-à-dire une approche scalaire basée sur l'idée d'un continuum.

1.1.1 Dichotomie : langue générale versus langue spécialisée

Généralement, la langue spécialisée est opposée à la langue générale. La langue générale est qualifiée comme la langue courante ou la langue commune, c'est-à-dire la langue quotidienne, qui est parlée, écrite et comprise par tous les locuteurs de la communauté linguistique et qui est utilisée dans la vie quotidienne. Elle se caractérise donc par l'acceptation générale et par sa large diffusion dans la communauté linguistique, tant à l'oral qu'à l'écrit. Signalons d'emblée que les études et les théories sur la polysémie et la sémantique portent surtout sur la langue générale et rarement sur la langue spécialisée. La langue spécialisée ou la langue de spécialité, par contre, s'emploie dans un domaine restreint, par exemple le domaine technique de la mécanique, de l'architecture, de l'électrotechnique, etc. Elle caractérise la communication scientifique des experts ou spécialistes d'une communauté restreinte dans le domaine spécialisé et permet de transmettre des informations particulières à ce domaine de connaissances.

Kocourek (1991a : 37-39) décrit la langue spécialisée à l'aide de trois aspects « de spécialité », à savoir (1) l'appartenance à un domaine donné, (2) l'intellectualisation et (3) la particularisation. L'appartenance à un domaine de spécialité correspond à une division horizontale ou thématique (par sujet), tandis que l'intellectualisation correspond à une division verticale ou stylistique. Le troisième aspect de spécialité est celui de la particularisation, c'est-à-dire l'accent mis sur le détail et les nuances. Selon Kocourek (1991a), la langue technique et scientifique accorde le plus d'importance à l'idéal de l'intellectualisation, parce qu'elle a tendance à définir les unités lexicales, à contrôler la polysémie et l'homonymie et à supprimer la synonymie. Les textes spécialisés sont caractérisés par des particularités lexicales, sémantiques, morphologiques et syntaxiques, notamment des termes spécifiques au domaine, des collocations particulières, des phrases plus longues, une surabondance de noms, de syntagmes nominaux et d'adjectifs et noms déverbaux, un suremploi de l'impersonnel et la présence prépondérante de déterminants plutôt que de pronoms.

Cabré (1998) définit les textes spécialisés par la concision, la précision et l'adéquation à la situation de communication. Ces trois aspects correspondent plus ou moins aux trois aspects de spécialité de Kocourek. Les termes jouent un rôle très important dans la langue spécialisée, car les termes (normalisés) servant à dénommer un concept spécialisé sont très concis. Ils contribuent à la précision et à la concision en évitant une paraphrase longue et complexe et permettent dès lors aux spécialistes de référer au domaine de spécialité de façon adéquate et efficace. Toutefois, pour tracer la frontière entre la langue générale et la langue spécialisée, il vaut mieux « recourir à des éléments extralinguistiques et communicationnels » (Cabré 1998 : 119), tels que le type d'interlocuteurs (spécialistes et semi-spécialistes), la situation de communication (discours spécialisé), les particularités stylistiques, etc. En effet, les unités linguistiques ne permettent pas toujours de bien distinguer la langue spécialisée de la langue générale (Cf. 1.1.2).

En allemand, la dichotomie entre langue générale et langue spécialisée se traduit par la dichotomie *Allgemeinsprache* ou *Gemeinsprache* (*Umgangssprache*) versus *Fachsprache* (Eriksen 2002 ; von Hahn 1998 ; Arntz & Picht 1989). La dénomination *Fachsprache* indique clairement le contexte spécialisé du métier (*Fach*). Selon Eriksen (2002), la *Fachsprache* est considérée comme la langue particulière de la spécialisation du métier ou du domaine d'activité. En anglais, la dichotomie se résume à l'opposition LGP (*Language for General Purposes*) versus LSP (*Language for Special Purposes*) (Bowker & Pearson 2002).

Avant de passer aux approches théoriques de la langue spécialisée et de la terminologie, il est intéressant de procéder à une explication « terminologique » concernant la différence entre la langue spécialisée et la langue de spécialité.

1.1.1.1 Langue spécialisée ou langue de spécialité ?

Force est de constater que l'usage est flottant. Certains auteurs préfèrent la dénomination *langue spécialisée*, notamment Lerat (1995b), Condamines et Rebeyrolle (1997), Condamines (1999), Dury (1999), Meyer & Mackintosh (2000), Van Campenhoudt (2000, 2001 et 2005). D'autres recourent à la dénomination *langue de spécialité*, comme Kocourek (1991a et 1991b), Lethuillier (1991), Gambier (1991), Gémar (1991), Delavigne & Bouveret (1999), Cabré (1998 et 2000a) et Sager (2000). Selon Habert et al. (1997), la *langue spécialisée* permet d'insister sur la continuité entre la langue générale et le fonctionnement particulier des usages spécialisés. Par contre, la *langue de spécialité* « met plutôt l'accent sur le domaine technique ou scientifique concerné » (Habert et al. 1997 : 148). Si elle est limitée au domaine des sciences et techniques, Kocourek (1991a et 1991b) mentionne la dénomination *langue technoscientifique* pour référer à la langue technique et scientifique. Gémar (1991) propose de parler de *technolecte*.

Il est à noter que l'Organisation internationale de normalisation ISO[1] préconise la dénomination *langue de spécialité*. En effet, la norme ISO1087-1 (1990), consacrée aux travaux terminologiques, définit la langue de spécialité comme un « sous-système linguistique qui utilise une terminologie et d'autres moyens linguistiques et qui vise la non-ambiguïté de la communication dans un domaine particulier » (ISO1087-1 cité dans Lerat 1995b : 17).

Par conséquent, il convient de situer la langue spécialisée par rapport à la terminologie. D'une part, la terminologie est considérée comme l'ensemble des unités terminologiques ou des termes, propres au domaine de spécialité. D'autre part, la terminologie est la science ou la théorie sous-jacente à l'étude des unités terminologiques, c'est-à-dire les principes et fondements conceptuels (Cabré 1998)[2], par exemple la terminologie traditionnelle ou la terminologie descriptive. Il est vrai que la langue spécialisée utilise principalement des unités terminologiques, mais elle ne se réduit pas à la terminologie au sens strict des unités terminologiques (Cf. 1.1.2.2).

Dans cette étude, nous adopterons la dénomination *langue spécialisée*, mettant en évidence la continuité entre la langue générale et les usages spécialisés. D'ailleurs, Lerat (1995b : 19) affirme qu'on ne saurait parler de langue de spécialité, car « il n'existe pas d'activités humaines entièrement cloisonnées ». Il préfère le participe passé *spécialisé*, en raison de la souplesse des interprétations : « il y a place pour des degrés variables de spécialisation, de normalisation et d'intégration d'éléments exogènes (soit empruntés, soit tirés de systèmes de signes non linguistiques insérés dans des énoncés en langue naturelle) » (Lerat 1995b : 20). Compte tenu de l'approche scalaire adoptée dans cette étude, il est clair que la dénomination *langue spécialisée* s'avère plus appropriée.

1.1.1.2 Langue spécialisée et terminologie traditionnelle

Une analyse de la langue spécialisée, et a fortiori du domaine technique de la machine-outil, serait incomplète sans une présentation des travaux d'Eugen Wüster, le père fondateur de la terminologie dite traditionnelle et l'auteur du premier « *Dictionnaire multilingue de la machine-outil* ». Un bref aperçu historique

---

[1] ISO = International Organisation for Standardization.

[2] Cabré (1998) distingue encore un troisième sens intermédiaire de la notion de terminologie, à savoir « l'ensemble des règles permettant de réaliser un travail terminographique » (nomenclatures).

permettra de mieux situer son approche de la langue spécialisée ainsi que la dichotomie entre langue générale et langue spécialisée.

La terminologie remonte à l'antiquité grecque, mais la terminologie moderne ne date que du début du XXième siècle. En 1906, la création de la Commission Electrotechnique Internationale (CEI) donne lieu au premier *Vocabulaire Electrotechnique International*. Ce que l'on appelle généralement la « Théorie classique de la terminologie » ou la « Théorie traditionnelle de la terminologie » et dès lors la « terminologie traditionnelle » (Cabré 2000a : 11), renvoie en fait à la *Théorie générale de terminologie* (TGT), conçue par Eugen Wüster (1898-1977) dans les années 1930. Ingénieur autrichien germanophone et spécialiste de la machine-outil et des vocabulaires spécialisés, il est particulièrement préoccupé par la précision et par l'efficacité de la communication spécialisée internationale. En publiant *Internationale Sprachnormung in der Technik : besonders in des Elektrotechnik*, une version étendue de sa thèse de doctorat[3], Wüster (1931) assoit les bases de la terminologie moderne et de l'École de Vienne.

Ce n'est qu'à partir des années 60 que l'on assiste à l'essor de la terminologie, grâce au développement des sciences et des techniques et à l'évolution des connaissances. Les spécialistes s'intéressent de plus en plus à la terminologie, poussés par la création de nouveaux concepts et par les besoins de dénomination qui en découlent. Comme spécialiste des glossaires multilingues, Wüster vise à mettre en place une communication scientifique et technique internationale efficace, par le biais de la normalisation et de la standardisation des unités terminologiques. En 1952, il prend en charge le comité technique TC 37, *Terminologie : principes et coordination*, fondé en 1936 et chargé d'élaborer les principes méthodologiques pour harmoniser les terminologies. Les publications les plus importantes de Wüster sont le *Dictionnaire multilingue de la machine-outil* (Wüster 1968) et la Théorie Générale de la Terminologie (de 1976), à savoir *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie* (Wüster 1991)[4].

Dans sa thèse, Wüster plaidait déjà pour « un dictionnaire où les termes sont organisés dans un ordre systématique, conformément aux relations notionnelles existant entre eux » (Wüster 1968 : xi) et c'est ce mode de classement qui est adopté dans son dictionnaire de la machine-outil (Wüster 1968). La terminologie de l'école de Vienne des années 60 et 70 se veut avant tout normative et prescriptive, visant à

[3] Titre en français : *La normalisation de la terminologie technique internationale* (1931).

[4] Ouvrage publié après la mort de l'auteur par son successeur et collaborateur Felber.

harmoniser les terminologies, c'est-à-dire les unités terminologiques des langues spécialisées.

La théorie générale de la terminologie ou la doctrine traditionnelle wüsterienne a pour objectif principal la « standardisation conceptuelle et dénominative » (Cabré 2000a : 12) de la communication professionnelle internationale, restreinte aux sciences et techniques. On part de l'identification et de l'établissement des concepts dans un champ de connaissances particulier pour en fixer les dénominations standardisées correspondantes (Cabré 2000a). La finalité visée, à savoir la précision et l'univocité de la communication professionnelle, sera abordée plus loin (Cf. 1.2). Afin d'atteindre son objectif de standardisation, la doctrine wüsterienne limite ses objets d'étude aux concepts et aux termes, les termes véhiculant les concepts en tant que dénominations linguistiques. Les concepts (ou notions) occupent une place centrale, permettant de caractériser la démarche wüsterienne comme une démarche proprement conceptuelle et onomasiologique. Le concept est internationalisable, parce qu'extralinguistique ou interlinguistique (Lerat 1995a), ce qui correspond bien aux préoccupations de la normalisation internationale. Le signifié linguistique n'est pas pris en compte. Le concept est antérieur ou préexistant à la dénomination (terme) et le terme est considéré comme étiquette du concept. Dès lors, la doctrine traditionnelle est aussi fortement référentielle.

Wüster est également l'un des principaux fondateurs de la normalisation terminologique et des normes ISO. Son ambition principale est l'amélioration de la communication internationale. Détail intéressant : Wüster croit à la terminologie (normalisée) comme il croit à l'espéranto (Gaudin 2005), la langue internationale conventionnelle. Etant donné que la réalité technique et scientifique change tous les jours, elle requiert toujours de nouvelles dénominations et des dénominations univoques. La normalisation s'avère donc indispensable pour freiner la multiplication éventuelle des dénominations et pour assurer « la prévisibilité, la sécurité et la qualité » (Lerat 1995b : 115). D'ailleurs, le libre développement de la langue technique entraîne « mauvaises formes, homonymie et synonymie » (Wüster 1931 : 131). Il est évident que la normalisation terminologique wüsterienne équivaut plutôt à la planification terminologique ayant pour seul objectif l'efficacité, la précision et l'univocité de la communication (internationale) scientifique et technique.

1.1.1.3 Dichotomie : mot versus terme

De ce qui précède, il ressort que la terminologie traditionnelle restreint son champ d'étude aux termes, c'est-à-dire aux dénominations des concepts ou des notions dans un domaine de connaissances spécialisées. Son approche prescriptive et normative onomasiologique préconise l'étude des unités terminologiques de la langue

spécialisée, qu'elle vise à standardiser et à imposer. D'où la dichotomie entre la langue générale (langue naturelle) et la langue spécialisée des sciences et techniques. La langue générale constitue l'objet de la lexicologie et de la lexicographie et se caractérise par une approche sémasiologique et descriptive, essentiellement linguistique. Le point de départ est la dénomination qui constitue l'entrée de dictionnaire et qu'on cherche à définir en regardant les contextes pour déterminer le(s) sens. La terminographie, par contre, étudie la langue spécialisée et ne regarde pas les textes spécialisés. Elle se situe en amont des textes spécialisés, car elle impose les dénominations en partant de la notion et en cherchant le terme approprié. L'approche de la terminographie est donc onomasiologique et prescriptive.

Au niveau des unités lexicales, cette dichotomie se traduit par la dichotomie entre mots et termes. Les mots font partie de la langue générale, tandis que les termes sont réservés aux langues spécialisées. Les termes sont généralement caractérisés comme des unités terminologiques simples ou complexes, linguistiques ou non linguistiques (contenant des chiffres, des signes, etc.) et utilisés dans un domaine spécialisé par des experts du domaine qui les définissent[5]. En fait, « les unités lexicales ne deviennent termes que si elles sont définies et employées dans les textes de spécialité » (Kocourek 1991a : 105). Les mots et les termes se distinguent seulement par leur mode de signification et par les conditions pragmatiques (Cabré 2000b). La signification des mots dépend « en grande partie de l'environnement linguistique », tandis que celle des termes « aurait été liée avant tout à l'environnement pragmatique » (Béjoint & Thoiron 2000 : 5). La dichotomie entre mot et terme (ou entre langue générale et langue spécialisée) caractérise l'approche traditionnelle catégorielle ou binaire des unités lexicales. Malheureusement, cette approche catégorielle est difficilement compatible avec l'étude de corpus spécialisés, comme le montrent les remises en question récentes, que nous détaillons ci-dessous.

1.1.2 Remises en question de la dichotomie

1.1.2.1 Terminologie descriptive et linguistique

Récemment, c'est-à-dire depuis l'essor des outils et des techniques de la linguistique de corpus et depuis la constitution de corpus électroniques de textes spécialisés, les adages de la terminologie traditionnelle ont été remis en question. Le développement de la linguistique de corpus et l'intérêt de la sociolinguistique et de la linguistique générale (Gaudin 2005), ont contribué à l'essor de la terminologie descriptive,

---

[5] Comme les termes sont définis par rapport à un domaine de connaissances spécialisées (externe à la *langue*), ils se situent au niveau de la *parole* (Kageura 2002).

résolument linguistique[6]. De même, la doctrine terminologique wüsterienne a fait l'objet d'une révision fondamentale. En particulier l'approche onomasiologique prescriptive et conceptuelle fut remise en question par les adeptes d'une approche sémasiologique descriptive et linguistique, basée sur l'étude de corpus de textes spécialisés. Citons notamment la « Théorie Communicative de la Terminologie » (Cabré 1998 et 2000a), la « Socioterminologie » (Gaudin 1993 et 2003) et la « Terminologie socio-cognitive » (Temmerman 1997 et 2000a) (Cf. 1.2.2.1). D'après la « Théorie Communicative de la Terminologie » de Cabré, outre que les connaissances générales et spécialisées ne peuvent pas être dissociées, les termes n'*appartiennent* pas de manière naturelle à un domaine déterminé, mais sont *utilisés* dans un domaine particulier. Dès lors, un terme pourra avoir des variantes dénominatives (même des synonymes), avec des valeurs pragmatiques égales ou différentes.

Même si la terminologie traditionnelle refuse toute approche linguistique de la terminologie, il est possible et envisageable de traiter l'ensemble des unités terminologiques à partir des théories linguistiques (Cabré 2000a). Les termes pourraient ainsi être décrits et considérés « comme des unités de forme et de contenu, dont l'usage dans certaines conditions discursives particulières leur fait acquérir une valeur spécialisée » (Cabré 2000a : 10). Bourigault et Slodzian (1999) plaident aussi pour un renouvellement théorique de la terminologie. Les constats empiriques et l'analyse de textes spécialisés les incitent à repenser les fondements théoriques de la terminologie, parce que « c'est dans le cadre d'une linguistique textuelle que doivent être posées les bases théoriques de la terminologie ». (Bourigault & Slodzian 1999 : 30). Kocourek (1991b) soutient également l'idée d'une approche descriptiviste et textuelle permettant l'étude du contexte (linguistique) des termes, notamment sous forme de collocations.

Si les partisans de la terminologie descriptive, résolument linguistique et textuelle, remettent en cause la dichotomie stricte entre langue générale et langue spécialisée, c'est surtout parce qu'elle ne tient pas suffisamment compte de la réalité langagière. Les termes font partie intégrante de la langue naturelle, tout en se caractérisant par le fait qu'ils véhiculent des connaissances spécialisées. La langue spécialisée est considérée comme « la langue elle-même, mais au service d'une fonction majeure : la transmission de connaissances » (Lerat 1995b : 21). De ce fait, les termes se caractérisent par une double attente : « il faut que ce soient des unités linguistiques intégrables dans des énoncés (…) et il faut en même temps que ce soient des unités

---

[6] La première description proprement linguistique des vocabulaires scientifiques et techniques est celle de 1982 de R. Kocourek (Cf. Kocourek (1991) pour une version étendue et révisée).

de connaissance à contenu stable, donc plus indépendantes du contexte que les mots ordinaires » (Lerat 1995b : 45). Ces deux caractéristiques correspondent aux facteurs contradictoires de flexibilité et de « systématicité »[7] (Kageura 2002). D'ailleurs, le degré de technicité et de spécialité de la langue spécialisée est variable et dépend des besoins de communication et du public visé, c'est-à-dire des variables de la situation de communication spécialisée. En plus, Lerat (1995b) insiste sur le fait que ce ne sont pas uniquement les termes qui véhiculent (ou qui dénomment linguistiquement) les connaissances spécialisées, mais qu'il y a également des emprunts ou des termes transcodés, tels que des sigles ou des symboles. La forte présence d'unités non linguistiques soulève d'ailleurs des questions sur les différents types d'unités linguistiques (ou unités lexicales) caractérisant la langue spécialisée et discutées ci-dessous (Cf. 1.1.2.2).

La remise en cause de la dichotomie entre langue générale et langue spécialisée et les analyses de corpus de textes spécialisés incitent aussi à réviser la notion de normalisation. A l'opposé de la planification terminologique wüsterienne, on ne pourra construire une *signification stable* pour les unités lexicales qu'à partir de leurs occurrences dans les textes spécialisés (Bourigault & Slodzian 1999). En effet, l'usage, qui se manifeste à travers les textes spécialisés authentiques, est le résultat des activités des spécialistes et reflète aussi leurs approches parfois différentes et concurrentes. Par conséquent, la description de l'usage, c'est-à-dire de la réalité langagière des textes spécialisés, devrait idéalement précéder ou accompagner l'effort normalisateur (Kocourek 1991b). Il ne s'agit donc aucunement de « nier l'intérêt ou la nécessité de la normalisation ». Au contraire, il faudra proposer une approche « qui s'appuie sur les réalités accessibles et analysables que constituent les textes spécialisés » (Béjoint & Thoiron 2000 : 15). De telle façon, le descriptif peut contribuer au prescriptif et à la rédaction de normes. Décrire pour mieux prescrire.

1.1.2.2 Unités lexicales de la langue spécialisée

Les approches descriptives linguistiques étudient et analysent la langue spécialisée à partir de corpus spécialisés. Certes, la langue spécialisée utilise principalement des termes propres au domaine, mais elle mobilise également « les ressources ordinaires de la langue » (Lerat 1995b : 21). « Technical texts, even those handled exclusively by experts, do not consistently use specialized technical vocabulary, nor does such vocabulary consist exclusively of established terms » (Opitz 1990 : 1058).

---

[7] Kageura (2002) adopte l'approche de la terminologie traditionnelle et ses travaux sont axés sur l'aspect rigide et systématique de la terminologie. Les partisans de la terminologie descriptive, par contre, préconisent la flexibilité de la terminologie.

En effet, un corpus de langue spécialisée, par exemple un corpus technique, ne contient pas uniquement des mots techniques ou « termes » au sens strict, propres au domaine spécialisé, tels que *usinage* ou *broche*, mais également des mots du VGOS ou du Vocabulaire Général d'Orientation Scientifique (Phal 1971). Ces mots s'emploient dans plusieurs domaines scientifiques et techniques et leur sens est déterminé par les contextes spécialisés (par exemple *machine*, *outil*). Finalement, le vocabulaire d'un corpus spécialisé comprend des unités linguistiques de la langue générale, tant des unités lexicales telles que *type*, *modèle*, *permettre*, que des unités grammaticales (prépositions, pronoms, etc.).

De même, Slodzian (2000) signale qu'entre les mots et les termes, il existe un « item tiers » dans les corpus spécialisés. C'est un terme non spécifique au sujet, faisant référence à un domaine externe, et il se situe « dans un continuum entre mot et terme » (Slodzian 2000 : 71). Cabré (1991) distingue également 3 couches de lexique du point de vue de la spécialisation, à savoir (1) le lexique général, (2) le lexique spécialisé ou lexique-charnière, c'est-à-dire le vocabulaire du tronc commun et (3) la terminologie proprement dite. La variabilité de la quantité de terminologie présente dans un texte dépend du degré d'abstraction et de technicité du texte (i.e. la situation de communication). Si le lexique-charnière est fréquent dans les textes spécialisés de large diffusion et de vulgarisation, la terminologie « représente le bloc restreint du vocabulaire utilisé par des spécialistes communiquant entre eux » (Cabré 1991 : 59). Une observation similaire se retrouve dans les textes du domaine juridique (Gémar 1991), qui se constituent autour d'un noyau dur de termes (la nomenclature). Ces termes sont associés à des cooccurrents précis du vocabulaire de soutien (vocabulaire quasi-juridique) et à des unités lexicales et grammaticales de la langue générale.

Ces études montrent que l'approche catégorielle et binaire (mot – terme), restreignant l'étude de la langue spécialisée aux seuls termes, n'est pas compatible avec l'étude de corpus de textes spécialisés (Cf. 1.1.3).

1.1.2.3 Interactions entre langue générale et langue spécialisée

Malgré tous les efforts de normalisation et de standardisation, les termes ne sont pas stables. Tout comme une langue formalisée recourt à la langue naturelle pour l'expression de formules mathématiques par exemple, une langue spécialisée ne pourra pas exister sans la langue générale, qui à son tour pourra tout de même parfaitement exister de façon autonome (Arnzt & Picht 1989).

En plus, les frontières entre langue générale et langue(s) spécialisée(s) sont floues (Delavigne & Bouveret 1999) et les interactions sont multiples. Il est parfaitement envisageable que les mots de la langue générale deviennent des termes (terminologisation), que les termes de la langue spécialisée deviennent des mots

(déterminologisation) et que les termes voyagent d'un domaine spécialisé ou technique à un autre (nomadisation). Ceci est une raison supplémentaire de remettre en cause la dichotomie (trop) stricte entre les mots de la langue générale et les termes de la langue spécialisée.

- *Terminologisation*

La terminologisation désigne le processus par lequel une forme linguistique connue (souvent un mot de la langue générale) devient un terme. Par métaphore, un mot pourra acquérir un nouveau contenu sémantique. Ainsi, des caractéristiques de personnes peuvent être attribuées à des machines (Arntz & Picht 1989), par exemple les *dents* d'une scie ou d'une roue. Les métaphores botaniques sont légion en mathématiques, par exemple *racine, sommet* (Pavel 1991). Rastier (1995) distingue quatre étapes de terminologisation ayant un effet d'objectivisation, à savoir la nominalisation, la lemmatisation, la décontextualisation et la constitution du mot en type. Ce qui est indispensable, c'est l'acquisition de traits particuliers. Une unité lexicale est terminologisée si elle acquiert une spécificité dans le domaine et « si elle n'est plus utilisée que dans des circonstances particulières et qu'elle ne peut plus servir de générique » (Sager 2000 : 52).

Toutefois, il semble que les critères de terminologisation invoqués couramment « ne sont ni nécessaires ni suffisants » (Lerat 1995b : 46). Lerat (1995b) mentionne quatre critères : premièrement, la présence d'un type de formant lexical (racine grecque, préfixe ou suffixe savant), facile à repérer par des analyses informatisées, malgré les contre-exemples, tels que *masse* en physique. Le second critère, le degré de figement syntaxique (par exemple l'absence de la préposition dans *imprimante laser*), constitue un « indicateur à utiliser avec esprit critique et culture » (Lerat 1995b : 47), parce que certaines expressions du français familier se caractérisent aussi par l'absence de déterminant ou de préposition (*promotion canapé, coin repas*). En plus, la variante avec préposition est attestée également (*imprimante au laser*), bien qu'elle soit moins fréquente. Ensuite, la proximité de paraphrases définitoires et de gloses explicatives constitue un critère classique, à manier également avec précaution. Finalement, Lerat invoque le critère le plus sûr, l'avis du spécialiste du domaine. Pour plus de certitude, Lerat propose le repérage d'un faisceau d'indices de terminologisation. Le fait que les critères de terminologisation invoqués ne sont pas toujours fiables et que les contre-exemples sont nombreux, corrobore l'idée de remise en question de la dichotomie entre mots et termes.

- *Déterminologisation*

Les interactions entre langue générale et langue spécialisée se manifestent également en sens inverse : les termes peuvent devenir des mots (Arntz & Picht 1989 ; Meyer & Mackintosch 2000). Les langues spécialisées influencent la langue générale, en

raison du rôle important des sciences et techniques dans la vie de tous les jours. Signalons à titre d'exemple le rôle grandissant de l'informatique. Des expressions spécialisées envahissent la langue générale et leur sens est parfois modifié par les non-spécialistes. La déterminologisation (par exemple *virtuel*) et la dilution (par exemple *stand-alone*) caractérisent l'infiltration de termes dans la langue générale.

La déterminologisation indique le fait que le sens terminologique (spécialisé) tend à s'élargir et que le terme repris par la langue générale adopte un sens plus général que lorsqu'il est utilisé dans un domaine spécialisé. Le noyau dur du sens est très général et le sens terminologique sous-jacent demeure quasi inchangé, en dépit d'un affaiblissement sémantique. Ainsi, le sens du mot *virtuel*, par exemple, est fortement lié au domaine de l'informatique (Meyer & Mackintosch 2000). Les mots déterminologisés subissent des changements conceptuels mineurs, notamment en raison de la compréhension superficielle par le non-spécialiste. Les ouvrages rédigés par des experts et destinés à des non-spécialistes « emploient les termes spécialisés de façon nettement moins rigoureuse » (Meyer & Mackintosch 2000 : 201). En plus, la compréhension superficielle des mots spécialisés influence aussi leur définition dans les dictionnaires de langue générale, qui proposent des définitions plutôt générales.

Le sens terminologique peut aussi être dilué, quand les mots déterminologisés se sont généralisés au point qu'ils ne désignent plus le concept d'origine. Il y a alors une perte de traits sémantiques et une rupture avec le domaine spécialisé d'origine. Par exemple, le mot *stand-alone* signifie dans le domaine informatique « ordinateur qui n'est pas relié à un réseau ». Mais dans la langue générale, *stand-alone* n'a plus aucun rapport avec les ordinateurs et renvoie à une « forme de statut indépendant », par exemple « *stand-alone stories, stand-alone toilet* » (Meyer & Mackintosch 2000 : 199). Lorsque le sens terminologique d'origine se dilue de manière significative pendant le processus de déterminologisation, il peut produire des usages familiers. Le changement sémantique s'accompagne dès lors de changements pragmatiques. Meyer et Mackintosch concluent que l'idéal se présente lorsque les unités lexicales susceptibles de se déterminologiser sont faciles à comprendre et à employer (p.ex. *souris, autoroute de l'information*). De la sorte, on évite une dilution trop importante.

La déterminologisation a évidemment des conséquences pour la langue générale, mais elle peut également affecter la langue spécialisée et même influencer la terminologisation. En effet, une unité lexicale déterminologisée qui est devenue un

mot peut, à son tour, ressurgir dans le domaine d'origine[8]. Toutefois, le sens est plus large que le sens terminologique d'origine, ce qui pourra donner lieu à une polysémie à l'intérieur du domaine. Le processus de déterminologisation peut ainsi « aboutir en une série de points situés sur une échelle allant de la langue terminologique la plus pointue à la langue très générale » (Meyer & Mackintosch 2000 : 212). La notion d'échelle est particulièrement intéressante, étant donné qu'elle rompt avec la vision traditionnelle de dichotomie ou d'opposition stricte entre mots et termes.

- *Nomadisation*

Les termes ne voyagent pas uniquement des langues spécialisées vers la langue générale, les migrations s'effectuent également entre deux ou plusieurs domaines spécialisés. En effet, les langues spécialisées ne sont pas hermétiques les unes aux autres et il faudrait les considérer « comme des territoires aux frontières perméables, plutôt que comme des univers clos » (Delavigne & Bouveret 1999 : 25). De nombreux domaines se recoupent ou sont devenus inter- ou multidisciplinaires, par exemple le domaine de l'environnement. Le fait que les termes voyagent d'un domaine spécialisé à un autre est qualifié de nomadisation, un processus qui concerne tant les termes appartenant à un domaine spécialisé, que les mots déterminologisés appartenant à la langue générale, mais venus d'un autre domaine spécialisé. Dans le dernier cas, la reprise du mot déterminologisé pourra donner lieu à une polysémie. La nomadisation affecte donc le sens des termes qui circulent et voyagent, mais il subsiste toujours un noyau de sens qui indique son origine (Gaudin 1993 ; Delavigne & Bouveret 1999). Ce noyau de sens est stable et n'est pas affecté par les transferts sémantiques du terme, car il se retrouve dans tous les domaines spécialisés impliqués dans la nomadisation. Cependant, des sens périphériques peuvent s'ajouter au noyau sémantique, ou disparaître en fonction de la circulation du terme dans les domaines spécialisés. En effet, hors des normes, des glossaires et des nomenclatures, les termes circulent : ils s'enrichissent et ils s'appauvrissent (Delavigne & Bouveret 1999). En raison de ces phénomènes de nomadisation, le décloisonnement de l'approche des vocabulaires spécialisés s'impose.

Comme les interactions entre langue générale et langue spécialisée et entre plusieurs langues spécialisées affectent le sens des termes et/ou des mots, les interactions remettent également en question la monosémie et l'univocité des termes, que nous expliciterons dans la section 1.2.2.

---

[8] Ainsi, le sens d'« immersion totale » de *virtuel* a disparu dans les *visites virtuelles* : « les visiteurs observent un lieu en trois dimensions sur un écran d'ordinateur, sans avoir recours à des lunettes spéciales ou d'autres instruments » (Meyer & Mackintosch 2000 : 212).

1.1.3 Solution alternative : un continuum

Il est clair que la dichotomie *mot – terme* ou *langue générale – langue spécialisée* est difficile à maintenir, compte tenu de la réalité langagière observée dans les corpus spécialisés. Remettant en cause l'opposition trop stricte entre langue spécialisée et non spécialisée, Delavigne et Bouveret (1999) proposent un continuum allant du texte de vulgarisation à la communication pointue entre professionnels très spécialisés. Nous nous proposons dès lors d'adopter l'idée de continuum comme solution alternative à la dichotomie traditionnelle *mot – terme*.

Dans notre étude, l'approche traditionnelle catégorielle ou binaire sera donc remplacée par une approche scalaire. Plutôt que d'opposer les mots aux termes, les unités lexicales du corpus spécialisé seront situées dans un continuum, plus particulièrement sur une échelle de spécificité (Cf. chapitre 4). L'échelle de spécificité sera plus facile à opérationnaliser que la dichotomie, du point de vue quantitatif du traitement automatique de la langue, en raison des degrés de spécificité calculés de façon automatique et objective.

1.2 SÉMANTIQUE ET POLYSÉMIE

Dans cette deuxième partie, nous abordons le modèle adopté pour l'analyse sémantique. Comme nous l'avons évoqué dans la partie précédente, la langue spécialisée se caractérise, idéalement, par la monosémie et par la monoréférentialité. La polysémie serait évitée grâce aux efforts de normalisation. Toutefois, des études récentes ont montré l'existence du phénomène linguistique de polysémie, même dans un domaine spécialisé. Par conséquent, nous proposons d'articuler cette partie autour de la dichotomie polysémie – monosémie (1.2.1) et de sa remise en question (1.2.2). La première section donnera d'abord des définitions plutôt générales de la monosémie, la polysémie, l'homonymie et l'indétermination[9] (1.2.1.1). Ensuite, elle abordera les approches sémantiques en linguistique (1.2.1.2) ainsi que l'approche monosémiste et homonymique de l'approche traditionnelle (1.2.1.3). La deuxième section sera consacrée à la remise en question de l'idéal de monosémie (1.2.2.1) et à celle des distinctions entre monosémie, polysémie, homonymie et indétermination (1.2.2.2). Nous passerons également en revue les études et expérimentations récentes sur la polysémie dans la langue spécialisée (1.2.2.3), tout en précisant la contribution que nous comptons apporter grâce à notre hypothèse alternative, axée sur l'idée d'un continuum et sur une approche quantitative (1.2.3).

[9] Le vague ou la sous-détermination (Kleiber 2004).

1.2.1 Dichotomie : polysémie versus monosémie

Avant d'aborder la dichotomie entre polysémie et monosémie, il convient de préciser les notions de base, en l'occurrence la monosémie, la polysémie, l'homonymie et l'indétermination. Signalons d'abord la différence d'emploi entre *signification*[10] et *sens*. La signification d'une unité linguistique se situe au niveau de la langue et a le statut de type (*type*), « constitué comme tel par le linguiste à partir des sens observés dans le discours » (Rastier 1994 : 34). La signification est le résultat d'un processus de décontextualisation. Le sens en revanche se situe au niveau de la parole, il a le statut d'occurrence (*token*) et il est actualisé en discours. En effet, le sens « suppose une contextualisation maximale » (Rastier 2003), aussi bien par le contexte linguistique, que par la situation.

1.2.1.1 Monosémie, polysémie, homonymie et indétermination

La monosémie caractérise les unités linguistiques qui n'ont qu'un seul sens : une forme exprime un sens et s'interprète de façon univoque. La polysémie, par contre, caractérise les unités linguistiques à plusieurs sens : une forme correspond à deux sens (bisémie) ou à plusieurs sens (polysémie). Les sens sont généralement apparentés ou reliés entre eux, par métaphore, par métonymie, par spécialisation (restriction de sens) ou par extension de sens. Les différents sens des unités polysémiques sont sémantiquement plus proches que les différents sens des unités homonymiques. L'homonymie explique le phénomène par lequel deux mots (étymologiquement) différents coïncident formellement. Un signifiant (une forme graphique ou sonore) correspond à deux ou plusieurs signifiés, mais il s'agit en fait de deux ou plusieurs signes différents. Généralement, les homonymes se caractérisent par des différences syntaxiques, par exemple un genre différent (*un tour – une tour*) ou une catégorie grammaticale différente (*le son – son chat*). Comme ce sont deux ou plusieurs mots différents, les signifiés ne sont pas reliés sémantiquement. Un quatrième cas de figure est celui de l'indétermination du sens (Fuchs 1996 ; Habert et al. 2005) ou de la sous-détermination (Kleiber 2004), c'est-à-dire du sens vague. Une unité linguistique est indéterminée ou sous-déterminée, si elle a un sens très général ou vague, qui est précisé ou enrichi par le contexte, par exemple *enfant* (« fille » ou « garçon ») ou *docteur* (« femme » ou « homme »)[11] (Kleiber 2004).

---

[10] En anglais : *meaning* (signification) versus *sense* (sens).

[11] Par exemple : *Notre docteur est enceinte.* (docteur : « femme »)

versus *Notre docteur a épousé une Tahitienne*. (docteur : « homme »).

Les critères permettant de différencier la polysémie de l'homonymie portent généralement sur la relation (souvent problématique) ou sur l'absence de relation entre les différents sens observés. Ce sont des critères synchroniques d'ordre sémantique. On avance également des critères étymologiques diachroniques et des critères morphologiques. Ainsi, des mots polysémiques sont censés présenter une cohérence sémique et avoir un étymon commun (*bureau* « table de travail » et « pièce où est installée la table de travail » (PR)). Des mots homonymiques, par contre, auraient des dérivés spécifiques selon les sens et des constructions syntaxiques différentes (Condamines & Rebeyrolle 1997). La différence entre la polysémie et l'homonymie est également explicitée au niveau de l'opposition langue versus parole (discours), c'est-à-dire en termes de significations et de sens. Ainsi, l'approche polysémique se caractériserait par la présence de plusieurs sens en discours (au niveau des occurrences), alors que l'approche homonymique consisterait à identifier plusieurs significations (types) au niveau de la langue, indépendamment des contextes d'usage (Cf. 1.2.1.3).

Pour opérer la distinction entre la polysémie et l'indétermination (ou le vague), on peut recourir à des tests logiques, linguistiques et définitionnels. Le test logique cadre dans l'approche véríconditionnelle : un mot est polysémique s'il est vrai ou faux simultanément pour le même référent (Geeraerts 1993) (Cf. l'indicateur d'autonomie sémantique de la « négation indépendante des lectures » (Kleiber 2004 : 9)). Il s'agit généralement d' « autohyponymes[12] » tels que *homme*. Le test linguistique ou le test classique de la contrainte d'identité concerne les restrictions sémantiques dans des phrases avec deux occurrences coordonnées d'un mot polysémique. La coordination à l'intérieur de la même phrase requiert leur identité sémantique, alors que les deux interprétations possibles sont une indication de la polysémie, par exemple pour le mot anglais *port*[13] (Cruse 1986 ; Geeraerts 1993). Le test définitionnel permet de distinguer plusieurs sens d'un mot, s'il n'y a pas une seule définition « minimalement spécifique » et « maximalement générale » (Geeraerts 1993 : 203). Une définition maximalement générale permettrait de couvrir l'extension totale du mot, c'est-à-dire tous les sens possibles. Une définition minimalement spécifique permettrait de distinguer le mot d'autres mots non synonymiques. Ainsi, pour le mot polysémique anglais *port*, il n'est pas possible de donner une seule définition maximalement générale (« entité »), couvrant tant le sens « harbour » que le sens « kind of wine », puisque cette définition ne permet pas

[12] Un autohyponyme désigne un mot qui présente, selon le contexte, une lecture hyperonymique ou générale et une lecture hyponymique (ou une interprétation de sous-catégorie), par exemple *homme* (« être humain » / « être humain mâle »).

[13] **At midnight the ship passed the port and so did the bartender.* (Geeraerts 1993 : 229).

de distinguer *port* d'autres mots (Geeraerts 2003). La pluralité de sens (autonomes) correspond à une pluralité de champs lexicaux (Kleiber 2004). Toutefois, signalons d'emblée que ces trois tests permettant de détecter la polysémie ne sont pas toujours fiables et génèrent parfois des résultats contradictoires (Cf. 1.2.2.2).

1.2.1.2 Approches sémantiques en linguistique

Rappelons que les études théoriques sur la monosémie et la polysémie portent quasi essentiellement sur la langue générale et rarement sur la langue spécialisée. Les exemples de polysémie cités dans la littérature sont d'ailleurs toujours des mots de la langue générale. En effet, la linguistique générale a tardé à se préoccuper de la terminologie (Gaudin 2005). Les travaux sur la terminologie, quant à eux, sont généralement consacrés à des questions ontologiques, à des nomenclatures et à l'extraction de termes. Rares sont les études sur la langue spécialisée consacrées à la sémantique ou à la polysémie : il est « difficile de trouver des travaux qui traitent explicitement de la polysémie dans la terminologie » (Condamines & Rebeyrolle 1997 : 177). Apparemment, les sémanticiens ne s'intéressent toujours pas à la sémantique de la langue spécialisée : « les sémanticiens continuent d'ignorer la terminologie, que l'on ne voit que rarement mentionnée dans les manuels ou traités de linguistique » (Béjoint & Thoiron 2000 : 15). Toutefois, force est de constater « qu'un nombre croissant de terminologues s'intéressent désormais à la linguistique » (ibid.) (Cf. 1.2.2.1) et qu'ils procèdent de plus en plus à l'étude de la polysémie dans la langue spécialisée (Cf. 1.2.2.3).

En linguistique générale, la polysémie a été étudiée selon différentes approches sémantiques, ce qui a donné lieu à une divergence des cadres théoriques. Généralement, on fait la distinction entre quatre grands courants théoriques dans la sémantique lexicale[14], mais nous n'entrerons pas dans les détails. Nous ne procéderons pas non plus à un survol historique de la sémantique. Nous nous contenterons en revanche de mentionner brièvement les approches sémantiques actuelles les plus courantes. Comme la polysémie est un phénomène omniprésent et dès lors incontournable dans la langue générale, de nombreuses études y ont été consacrées. « La polysémie est un casse-tête chinois pour toute théorie sémantique » (Kleiber 2002 : 89). Si du moins l'on est d'accord sur la définition générale de la polysémie (pluralité de sens apparentés), les explications théoriques diffèrent, car elles se basent sur des prises de position quant aux problèmes inhérents à toute approche du sens (Kleiber 2002). Faut-il voir la polysémie « sous l'angle de la discontinuité sémantique (sens discrets) ou de la continuité sémantique (caractère

[14] A savoir la sémantique préstructuraliste, la sémantique structuraliste et néostructuraliste, la sémantique générativiste et néogénérativiste et la sémantique cognitiviste (Geeraerts 2002).

graduel et continu des sens polysémiques) » (Kleiber 2002 : 91) ? Kleiber (2002) soulève également le problème de l'apparentement des sens multiples : « faut-il un sens schématique supérieur ou non ? » ou « faut-il postuler un sens de base duquel dérivent les autres ? » (Kleiber 2002 : 91).

Ces questions relatives à la polysémie en général nous amènent aux approches concernant la polysémie en français, à savoir la sémantique référentielle et cognitive (Kleiber), la sémantique componentielle et différentielle interprétative (Rastier), la sémantique de la construction dynamique du sens ou le constructivisme modéré (Victorri et Fuchs) et le constructivisme radical et son modèle génératif (Récanati).

- *La sémantique référentielle et cognitive*

Les trois lignes de force de l'approche de Kleiber (1999) sont : (1) une sémantique référentielle dans un cadre réaliste et positiviste et tournée vers la réalité, (2) une sémantique à vocation cognitive et ancrée dans l'expérience humaine, (3) le maintien d'un sens « linguistique », conventionnel, « a priori ou préconstruit » et « intersubjectivement stable » (Kleiber 1999 : 12). La dernière caractéristique soulève la question de la nécessité du sens référentiel[15] et conventionnel, remis en question par le constructivisme radical (Cf. Récanati ; Kayser) et par le constructivisme modéré (Cf. Victorri & Fuchs).

Signalons à cet effet la polémique qui s'est engagée entre 1987 et 1991 dans les revues *Langages* et *Linguisticae Investigationes*. Kayser, informaticien et constructiviste, suggère l'idée d'une « sémantique qui n'a pas de sens » (Kayser 1987 : 33) et plaide pour la primauté des règles d'inférence sur le sens. Pour le mot *livre*, il serait ainsi possible de rendre compte de la multiplicité de types de référents possibles sans recourir à la notion de sens référentiel. Kleiber et Riegel répondent notamment par le principe de métonymie intégrée : « Certaines caractéristiques de certaines parties peuvent caractériser le tout » (Kleiber & Riegel 1989 : 414). Le principe de métonymie intégrée « permet de réguler de façon satisfaisante la variation référentielle en interaction prédicative sans multiplier inutilement les cas de polysémie » (Kleiber 1999 : 86) et permet aussi d'expliquer « des phénomènes référentiels sans postuler un changement de référent » (ibid. : 145). Notons dans ce

---

[15] « Le sens référentiel ou dénotatif ou encore vériconditionnel est ainsi un faisceau de traits intrinsèques ou inhérents du référent, ou encore traits objectifs, c'est-à-dire des traits qui sont supposés être possédés par le référent, donc des traits référentiels, en lien avec la réalité » (Kleiber 1999 : 35).

contexte la solution de Pustejovski (1995) par rapport à la multiplicité des sens lexicaux : la coercition de type[16] ou le changement de type.

Kleiber situe la question de la polysémie dans la version étendue de la sémantique du prototype (sémantique cognitive) : « l'unité lexicale polysémique constitue elle-même une catégorie » et « l'apparentement des sens multiples réside dans une organisation du type ressemblance de famille » (Kleiber 2002 : 94).

- *La sémantique componentielle et différentielle interprétative*

Rastier (1994) reproche au lexique génératif de Pustejovsky de s'inscrire dans une perspective fortement référentielle, en dépit de son « objectif de définir un formalisme général permettant une décomposition lexicale » (Rastier 1994 : 97). La sémantique componentielle et différentielle de Rastier se distingue de la sémantique référentielle, parce qu'elle « permet d'opposer deux formes linguistiques par un ou plusieurs traits de sens » (Normand 1999 : 121), ce qui la rend *différentielle*. Elle est *componentielle*, parce qu'elle décompose la signification en plusieurs traits de sens ou sèmes. On retrouve donc l'idée structuraliste de compositionnalité. Rastier oppose les sèmes inhérents (propriétés inhérentes héritées du type) aux sèmes

---

[16] Le modèle génératif de Pustejovsky (1995) permet de « rendre compte de la multiplicité des sens lexicaux rencontrés sans recourir à une approche du type *word sense enumeration* » (Kleiber 1999 : 185).

Selon le modèle génératif (Pustejovsky 1995 ; Pustejovsky & Boguraev 1996), chaque unité lexicale se caractérise par une représentation sémantique, comprenant quatre niveaux ou structures : (1) une structure argumentative (qui précise le nombre et le type d'arguments d'une unité lexicale), (2) une structure événementielle (qui identifie le type d'événements : un état, un processus ou une transition), (3) une structure *qualia* (qui définit les aspects essentiels de la signification des objets) et (4) une structure d'héritage (qui précise comment les unités lexicales sont reliées) (Pustejovsky 1995). La structure *qualia* comprend quatre rôles *qualia*, à savoir le rôle constitutif (matériel, poids, etc.), le rôle formel (dimension, orientation, position, couleur, etc.), le rôle télique (but ou fonction) et le rôle agentif (origine). Ces quatre niveaux de représentation sémantique sont connectés par des mécanismes génératifs généraux, tels que la coercition de type, permettant l'interprétation des mots en contexte.

Citons en guise d'exemple le syntagme verbal *commencer un livre*. Le verbe *commencer* porte toujours sur un événement, alors que *livre* est défini comme un objet physique dans sa structure argumentative. Le mécanisme de coercition de type prévoit que le verbe *commencer* impose son propre type sémantique (événement) à son argument *livre*. Celui-ci change de type pour passer d'un objet physique à un événement. Ce changement de type est possible pour *livre*, parce que la structure *qualia* du mot *livre* comprend un rôle télique qui permet au livre d'être lu (événement) et un rôle agentif qui permet au livre d'être écrit (événement). Ainsi, le type sémantique est respecté (événement), mais la syntaxe de l'expression ne change pas.

afférents (« traits sémantiques dont l'actualisation résulte d'une contrainte contextuelle » (Rastier 1994 : 38)). En faisant intervenir le contexte dans la construction, parce qu'il permet de dégager les traits de sens, la sémantique différentielle s'inscrit dans une perspective *interprétative* et textuelle : le texte détermine le sens des mots, à partir de leur signification en langue, mais en l'enrichissant.

- *La construction dynamique du sens ou le constructivisme modéré*

Comme l'idée de l'unicité (sémantique) du mot est plus importante pour distinguer entre homonymie et polysémie que le critère étymologique, Victorri et Fuchs (1996) recourent à des critères sémantiques pour définir la polysémie. Serait-il possible de trouver « des éléments de sens communs entre les différentes acceptions » ou « des sens intermédiaires entre les emplois les plus éloignés » (Victorri & Fuchs 1996 : 12) ? Ils rejettent l'idée de listes exhaustives de sens potentiels préétablis et, dans une perspective de sémantique dynamique, ils plaident pour la construction dynamique du sens, établi ou construit en interaction avec les éléments linguistiques et extralinguistiques du contexte (Victorri & Fuchs 1992 et 1996 ; Victorri 1997a et 1997b).

> Le sens d'un énoncé est le résultat d'un double mouvement, puisque ce sens est évidemment fonction du sens des expressions qui le composent, mais qu'inversement le sens de ces expressions dans cet énoncé est fonction du sens global de l'énoncé lui-même. (Victorri & Fuchs 1996 : 41)

Le calcul du sens entendu comme la construction dynamique du sens s'inscrit dans la *Gestalttheorie* : « le tout est plus que la somme de ses parties » et « une partie dans un tout est autre chose que cette partie isolée ou dans un autre tout » (Victorri & Fuchs 1996 : 41). Ainsi, le calcul du sens est un « processus dynamique au cours duquel les sens des différents mots s'influencent mutullement et qui aboutit simultanément à la détermination du sens de chacun des mots et à un sens global pour la phrase » (Venant 2004 : 1147). Par conséquent, le sens d'une unité lexicale polysémique « peut se définir et s'analyser, par des méthodes linguistiques, à partir des relations qu'elle entretient dans les différents systèmes paradigmatiques et syntagmatiques auxquels elle prend part » (Victorri & Fuchs 1996 : 199). La construction dynamique du sens, ou le principe de la compositionnalité « gestaltiste », maintient l'idée de sens linguistique associé au mot, c'est-à-dire un noyau de sens invariant (Victorri & Fuchs 1992), une sorte de « sens schématico-dynamique », qui est complété par l'interaction avec le contexte. La polysémie est définie comme « la trace, dans le système de la langue, d'un processus qui va de la

forme schématique instable à l'infinité des effets de sens distincts dans les conditions toujours spécifiques de la parole » (Victorri 1997a : 59).

- *Le constructivisme radical*

L'approche sémantique de Récanati (1997) s'inscrit également dans la perspective contextuelle, mais dans une conception générativiste et dans un constructivisme plus radical. Ce qui unit le constructivisme modéré de Victorri et Fuchs et celui de Récanati, plus radical, est l'idée que les énoncés n'ont pas de « conditions de satisfaction en vertu purement de leur signification linguistique » (Récanati 1997 : 120). Mais Récanati abandonne entièrement l'idée d'un sens linguistique fixe. Le sens des mots « n'est pas fixé une fois pour toutes ». En plus, « la variation en question n'est pas diachronique mais synchronique : même relativement à un état de langue donné, le sens des mots varie systématiquement d'une occurrence à l'autre » (Récanati 1997 : 107).

Il propose de recourir à un modèle génératif, non pas pour « sélectionner le sens pertinent dans une liste de sens possibles préétablis », mais pour « engendrer le sens pertinent » (Récanati 1997 : 114). Le modèle génératif devrait permettre de rendre compte du nombre infini de sens potentiels ainsi que du caractère graduel et continu des sens, contrairement aux sens discrets de l'approche traditionnelle et fixiste. En effet, « adopter un tel modèle génératif revient à admettre la variabilité contextuelle du sens et donc à abandonner le fixisme » (ibid.).

1.2.1.3 Terminologie traditionnelle « monosémiste » et homonymique

Après ce bref résumé des approches récentes de la polysémie, revenons à la dichotomie entre polysémie et monosémie, afin d'expliquer pourquoi la terminologie traditionnelle préconise la monosémie et exclut la polysémie.

Comme nous l'avons mentionné dans la première partie, la terminologie traditionnelle et les efforts normalisateurs de Wüster visaient principalement la précision et l'efficacité de la communication professionnelle entre les spécialistes du domaine. Les besoins communicatifs dans la langue spécialisée requièrent plus de précision, ce que la terminologie traditionnelle identifie et définit comme le principe de la bi-univocité (*Eineindeutigkeit*), à savoir la monosémie et l'univocité : chaque concept est désigné par un terme et chaque terme dénomme un concept (Wüster 1931 et 1991). La terminologie traditionnelle « conceptuelle » accorde une importance capitale au concept ou à la notion, car il est le point de départ de la terminologie. La bi-univocité entre la notion (sens) et la dénomination (forme) et entre la dénomination et la notion implique que l'homonymie (*Mehrdeutigkeit*) et la synonymie sont évitées ou limitées (Wüster 1931 : 94). Dans les travaux de Wüster, il est aussi rarement question de polysémie. La terminologie traditionnelle préconise

donc pour les termes de la langue spécialisée, la monoréférentialité (chaque terme a un référent) et la monosémie (chaque terme a un sens). Ainsi, Wüster désire « surmonter les difficultés de la communication professionnelle, difficultés qui trouvent leur origine, selon lui, dans l'imprécision, la diversification et la polysémie de la langue naturelle » (Cabré 2000a : 11). C'est en raison de cet idéal de monosémie dans les textes spécialisés que les partisans de la terminologie traditionnelle (notamment Wüster et son successeur Felber) sont souvent qualifiés de « monosémistes ».

Pour le français, les pioniers de la terminologie ont été Rondeau et Guilbert. Guilbert (1973) plaide pour la monosémie et la monoréférentialité des langues spécialisées, en insistant sur l'appartenance au domaine spécialisé, caractéristique principale et définitoire des unités terminologiques.

> Le terme technique tend à être monosémique ou plutôt monoréférentiel dans chaque domaine particulier de la connaissance. Les choses du monde, qui sont perçues et comprises par leurs éléments essentiels, doivent être classées et distinguées ; les termes techniques et scientifiques qui les désignent, pour éviter l'ambiguïté et la confusion dans la communication, ne désignent qu'une chose. C'est pourquoi chaque vocabulaire technique et scientifique forme un ensemble dont les éléments sont structurés du fait même de leur appartenance à un vocabulaire et non à un autre, le terme n'y figure que par sa référence à ce domaine particulier. (Guilbert 1973 : 11)

Il convient de se pencher également sur l'approche homonymique de la terminologie traditionnelle, qui résulte de l'idéal de monosémie. Si le même terme s'emploie dans deux ou plusieurs domaines différents, il n'est pas considéré comme potentiellement polysémique, ayant deux ou plusieurs sens spécialisés. Selon les monosémistes, il s'agit de deux termes homonymiques, car employés et définis dans des domaines différents. D'ailleurs, Wüster (1931) signale qu'on ne peut prétendre à l'univocité absolue. Il suffit que les termes soient univoques en contexte et à l'intérieur du domaine spécialisé. Ainsi, la même forme peut revêtir des sens différents dans des domaines différents, puisqu'un domaine de spécialité est un champ fermé. La primauté du concept sur la dénomination et l'optique résolument référentielle (Gaudin 1995b) entraînent donc la multiplication des homonymes. L'approche homonymique permet d'expliquer l'existence de plusieurs sens différents, dans plusieurs domaines spécialisés, et de maintenir ainsi le principe de la monosémie et de la précision à l'intérieur du domaine spécialisé.

1.2.2 Remises en question de la dichotomie

1.2.2.1 Terminologie descriptive et linguistique

La terminologie descriptive, linguistique et textuelle, s'inscrit dans une perspective sémasiologique et elle remet naturellement en question l'idéal d'univocité et de monosémie préconisé par la terminologie traditionnelle[17]. Dans cette perspective, l'analyse de corpus de textes spécialisés permet d'attester la polysémie et la synonymie (variantes lexicales), même dans la langue spécialisée et même à l'intérieur d'un domaine spécialisé.

Dans sa « Théorie Communicative de la Terminologie » (TCT), Cabré (1998) émet une réserve par rapport à l'idéal d'univocité, signalant que les langues de spécialité « tentent (mais tentent seulement) de disposer d'une dénomination pour chaque concept, tout en tolérant, dans une certaine mesure, la synonymie » (Cabré 1998 : 117). Elle émet également une réserve par rapport à l'idéal de monosémie, soutenant que les langues de spécialité « n'ont pas, *en théorie*, de termes polysémiques » (ibid.), puisque la polysémie du lexique commun devient l'homonymie dans le lexique spécialisé. Si l'importance de la polysémie et de la synonymie est souvent sous-estimée ou négligée, c'est en raison des normes, de la terminologie et des nomenclatures (Cabré 1991). Mais en fait, les « termes réels sont potentiellement polysémiques, parce que leur signifié peut être élargi et multiplié dans différents domaines de spécialité » (Cabré 2000b : 35). Les dénominations utilisées dans deux ou plusieurs domaines étant formellement identiques, elles relèvent de la même unité lexicale, qui est dès lors polysémique. Utilisée dans deux ou plusieurs domaines, l'unité lexicale polysémique revêt soit le même sens, soit des sens différents, mais qui sont tirés de la même unité de base (par exemple *virus* dans le domaine de la médecine et *virus* dans le domaine de l'informatique).

De nos jours, on ne peut plus nier le caractère interdisciplinaire des termes. Cependant, les termes reçoivent une seule définition dans un vocabulaire défini et précis. Si les unités formelles s'emploient dans plusieurs domaines, elles sont reprises dans autant de dictionnaires spécialisés, avec une seule définition précise par domaine. Toutefois, il est clair que les définitions présentent une similitude, malgré « la séparation physique et sémantique » (Cabré 2000b : 32). Cabré considère que toutes les unités lexicales sont polysémiques, car « la polysémie implique le fait d'être associé à des groupes de traits sémantiques qui s'activent selon les différentes situations » (Cabré 2000b : 34). Même si quelques unités lexicales sont « associées momentanément à un seul sens et utilisées dans un

[17] Théorie Générale de la Terminologie (Wüster 1991).

domaine de spécialité », elles sont tout de même susceptibles d'« incorporer un nouveau sens quand elles sont utilisées dans un domaine thématique différent » (ibid.) (Cf. nomadisation des termes). En raison de l'idéalisation de la connaissance spécialisée, antérieure à toute expression et uniforme dans toutes les langues, l'approche traditionnelle est incapable d'expliquer l'interdisciplinarité. D'ailleurs, en se limitant à la standardisation et à la normalisation, elle ne rend aucunement compte des données empiriques, ni de la réalité langagière de la communication spécialisée. Pourtant, les observations empiriques démontrent la portée limitée de la normalisation et de l'approche prescriptive et permettent de relever dans les textes spécialisés des preuves de variabilité terminologique et de polysémie, notamment en raison de l'interdisciplinarité croissante des technologies (Slodzian 2000).

La « Socioterminologie » de Gaudin (1993 et 2003) remet également en question l'univocité et la monosémie, en adoptant une approche sociolinguistique et descriptive de la terminologie à partir de l'exploration de la vulgarisation scientifique. Il lance l'idée que la polysémie fait « *boule de neige* » et que « le succès de certains termes pousse à leur reprise » (Gaudin 1993 : 107), parce que toute énonciation s'inscrit dans un ensemble de discours « énonçables » et acceptables. Il insiste aussi sur le fait que la métaphore tisse des liens entre la langue de la recherche et la langue commune. Gaudin questionne également l'approche synchronique structuraliste et homonymique de la terminologie traditionnelle. Les domaines de spécialité ne sont pas nettement délimités (*clear-cut*). Gaudin rejette l'idée d'une appartenance exclusive à un domaine et propose l'idée d'un continuum entre science et technique et le « fonctionnement dans le cadre d'une activité » (1993 : 83). Il est clair que le même terme peut recouvrir plusieurs notions dans des domaines différents, mais il convient de se demander « si derrière les concepts nomades ne sont pas véhiculés des quasi mêmes notions » (Gaudin 1993 : 109).

Temmerman (1997, 2000a et 2000b) plaide pour la « Terminologie socio-cognitive » et elle rejette l'approche synchronique et prescriptive wüsterienne en faveur d'une approche diachronique et descriptive, mettant en évidence la fonctionnalité de la polysémie et de la synonymie dans la langue spécialisée. Elle insiste surtout sur les aspects conceptuels de la polysémie et de la synonymie. Du point de vue sémasiologique et diachronique, la polysémie est le résultat de la réflexion humaine sur le monde, c'est-à-dire le résultat synchronique de l'évolution sémantique, qui est un accroissement diachronique de la densité d'informations (Temmerman 2000a). « Au lieu de partir de la notion clairement délimitée, la

terminologie sociocognitive part des unités de compréhension[18], caractérisées le plus souvent par une structure prototypique » (Temmerman 2000b : 59). Temmerman (2000a et 2000b) soulève trois causes potentielles de polysémie : (1) l'évolution des unités de compréhension ; (2) la flexibilité et l'adaptation des catégories de structure prototypique[19] en raison de l'innovation technologique ou sociologique (la perception) et (3) l'adaptation des moyens d'expression, c'est-à-dire la dynamique de la langue. Par conséquent, la polysémie et la synonymie contribuent à la compréhension, à la perception et à l'expression des connaissances spécialisées (Temmerman 2000a et 2000b).

La terminologie descriptive linguistique, sémasiologique et textuelle adopte une méthodologie distributionnelle et contextuelle. Le terminologue étudie désormais la distribution des unités lexicales : il analyse leur contexte linguistique et communicatif réel afin d'identifier le(s) sens (différents). L'étude systématique des cooccurrences des unités lexicales s'avère indispensable pour les définir. Cette approche distributionnelle et contextuelle de la terminologie descriptive s'oppose à l'approche référentielle de la terminologie traditionnelle et prescriptive, où les unités lexicales sont traitées de façon isolée et où l'axe syntagmatique n'est aucunement pris en compte pour la désambiguïsation.

> En raison de cette monosémie référentielle, inhérente au terme lui-même, à l'opposé de ce qui se passe pour le terme du lexique général, l'axe syntagmatique de la phrase n'intervient pas pour lever une ambiguïté éventuelle du nom dans la communication entre spécialistes. (Guilbert 1973 : 11)

Quelques expérimentations récentes menées sur des corpus spécialisés s'inscrivent dans cette perspective distributionnelle et contextuelle de la terminologie descriptive (Cf. 1.2.2.3). Comme le contexte linguistique d'une unité polysémique permet de la désambiguïser et donc de choisir le sens pertinent, l'axe syntagmatique, et plus précisément les cooccurrents et les collocations, se révèlent indispensables pour l'analyse sémantique de l'unité lexicale (Cf. chapitre 5).

---

[18] Ce sont des « *units of understanding* » (Temmerman 2000a : 153). « Nous utilisons le terme d'*unité de compréhension* pour désigner les catégories de structure prototypique et pour les notions clairement délimitables » (Temmerman 2000b : 59).

[19] Les variantes peuvent être incorporées dans une catégorie en raison de la ressemblance avec le prototype.

1.2.2.2 Remise en question des critères de distinction

Les mots de la langue générale sont dits avoir un ou plusieurs sens : ils sont monosémiques ou polysémiques. Cette dichotomie mérite quelques mises au point, parce que la réalité langagière n'est pas si simple et transparente qu'elle puisse être appréhendée en termes de dichotomie.

> In most accounts of contextual variation in the meanings of a word, a sharp distinction is drawn between « one meaning » and « many meanings », between monosemy and polysemy. But this is too crude : there are many degrees of distinctness which fall short of full sensehood, but which are none the less to be distinguished from contextual modulation. (Cruse 2000 : 114)

A l'encontre de la polysémie et de l'homonymie qui sont toutes les deux qualifiées de « plurivocité » (un signifiant ayant plusieurs signifiés), la monosémie et l'indétermination (ou la sous-détermination) attachent un seul signifié (même s'il est vague) à un signifiant et relèvent donc de l'univocité. Or, cette subdivision est discutable parce que trop stricte, comme l'est d'ailleurs la distinction entre polysémie et homonymie et entre polysémie et indétermination.

En ce qui concerne la « plurivocité » ou la pluralité des sens, un critère sémantico-paradigmatique permettrait de distinguer entre polysèmes et homonymes (Cf. 1.2.1.1). Premièrement, la polysémie suppose une cohérence sémique, contrairement à l'homonymie, car les différents sens d'un mot polysémique sont sémantiquement reliés. Deuxièmement, les polysèmes auraient des synonymes et des antonymes identiques, contrairement aux homonymes (Condamines & Rebeyrolle 1997). Toutefois, ce critère sémantico-paradigmatique n'est pas fiable, ce qui se reflète également dans les dictionnaires, plus particulièrement par « des écarts sensibles dans la répartition des polysèmes et des homonymes d'un dictionnaire à l'autre » (Condamines & Rebeyrolle 1997 : 175). En outre, certains lexicographes distinguent les sens différents d'une unité lexicale « plurivoque » comme appartenant à autant d'homonymes différents, ce qui se traduit par des entrées différentes. D'autres lexicographes considèrent ces sens différents ou quelques-uns de ces sens comme étant reliés sémantiquement et ils les regroupent par conséquent sous la même entrée polysémique[20]. Mais sous celle-ci, on ne retrouve pas partout le même nombre de

---

[20] « Dans de nombreux cas, des *lexies* ont un même signifiant et, en plus, manifestent entre elles des liens sémantiques assez évidents » (Mel'čuk et al. 1995 : 155). « Les lexies montrant la relation de polysémie entre elles seront regroupées en des ensembles appelés *vocables* » (Mel'čuk et al. 1995 : 15). Un vocable correspond à un article de dictionnaire d'un mot polysémique dans les dictionnaires courants.

sens recensés, ni les mêmes distinctions sémantiques, données sous forme de définition. Par conséquent, le critère de la cohérence sémique ne conduit pas toujours à des résultats convergents, ni en termes de regroupement de sens recensés, ni en termes paradigmatiques (synonymes et antonymes).

Notons que l'interprétation de la notion d'« *ambiguïté* » manque aussi de clarté. L'ambiguïté implique nécessairement un choix, car les sens différents (ou les significations différentes) sont distincts et mutuellement exclusifs (Fuchs 1996). Fuchs considère tant les polysèmes que les homonymes comme étant ambigus. Dans d'autres études (e.a. Cruse 1986 ; Tuggy 1993 ; Geeraerts 1993), l'ambiguïté porte uniquement sur les homonymes. Selon Nerlich et al. (2003), il existe même un cycle dans le temps ou un processus continu, se constituant (1) de polysémie émergente[21], (2) de polysémie conventionnalisée et lexicalisée[22] et (3) de polysémie morte ou d'homonymie[23]. L'idée d'un continuum se retrouve chez Victorri (1997a)[24] et chez Klepousniotou (2002). Klepousniotou (2002) étudie le traitement mental de l'ambiguïté lexicale. S'il est vrai que les sens reliés ou non reliés sémantiquement permettent généralement de distinguer entre polysémie et homonymie, Klepousniotou propose de considérer un continuum allant de la polysémie « pure » à l'homonymie « pure » selon le degré de parenté sémantique. Ainsi, la polysémie métaphorique, qui s'appuie sur une relation d'analogie, se trouverait plus près de l'homonymie, tandis que la polysémie métonymique, basée sur une relation de contiguïté, se situerait à l'autre bout du continuum.

D'ailleurs, force est de constater que les critères synchroniques (cohérence sémique des polysèmes) et diachroniques (étymon identique des polysèmes) de la distinction entre polysémie et homonymie ne sont pas toujours convergents. Ainsi, les locuteurs

---

[21] Nous entendons par « polysémie émergente » un emploi métaphorique (par analogie) ou un emploi métonymique (par contiguïté) qui n'est pas encore conventionnalisé. Ce n'est pas un cas d'ambiguïté, mais plutôt d'indétermination (Victorri & Fuchs 1996).

[22] C'est le cas quand les différents sens sont recensés dans le dictionnaire (p.ex. *bureau*).

[23] C'est le cas quand les mots sont considérés comme des homonymes, en dépit de l'étymon commun (Cf. ci-dessous *voler* et *grève*).

[24] Entre la polysémie et l'homonymie, il y a « un véritable continuum qui joue un rôle important dans l'évolution de la langue et qui rend impossible d'effectuer en synchronie une dichotomie pure et simple » (Victorri 1997a : 57). En diachronie, on peut observer de « lentes dérives de la polysémie vers l'homonymie » (Victorri 1997a : 60).

considèrent normalement qu'il existe trois sens de *bureau*[25], mais l'étymon est commun. En témoigne l'entrée unique et polysémique de *bureau* dans la plupart des dictionnaires. Fuchs (1996) donne également l'exemple de *voler* « dérober » et *voler* « se déplacer dans l'air au moyen d'ailes », où la langue « semble avoir perdu le souvenir d'une étymologie commune » (Fuchs 1996 : 27). En effet, dans la plupart des dictionnaires, les deux verbes *voler* sont considérés et présentés comme des homonymes. Toutefois, la signification « dérober » procède du verbe *voler* « se déplacer … », « utilisé en emploi transitif à propos du faucon qui attaque sa proie » (Fuchs 1996 : 27). Il en va de même pour le substantif *grève*[26] (« terrain plat au bord d'une rivière » et « arrêt collectif d'un travail ou d'une activité »). Dans ces exemples, comme dans d'autres, la polysémie se trouve transformée en homonymie.

En linguistique, il existe deux courants qui nient tous les deux la spécificité de la polysémie, à savoir la stratégie de la polysémie « réduite » et celle de la polysémie « éclatée » (Fuchs 1996). La polysémie réduite revient à ramener des cas d'ambiguïté à des cas d'indétermination ou de sous-détermination. Elle consiste à réduire la polysémie à une sorte « d'univocité sous-déterminée » (Fuchs 1996 : 32) basée sur un noyau de sens unique et sous-déterminé. Le courant de la polysémie réduite s'inspire du structuralisme et de son principe de bi-univocité des rapports entre forme et sens. A chaque mot (forme) correspond un seul noyau de sens en langue, que le contexte est censé déterminer et enrichir. Ce noyau de sens constitue la valeur lexicale de l'expression, tandis que tout autre sens en sera une valeur contextuelle (Bianchi 1991). Les significations sont donc évacuées hors de la langue et considérées comme « surdéterminées » (ou qualifiées d'« effets de sens en discours » ou de « significations référentielles extralinguistiques ») (Fuchs 1996 : 32). Le noyau de sens unique en langue et largement sous-déterminé est censé être sous-jacent à la diversité de significations en contexte. La variation se trouve ainsi conditionnée par le contexte. L'idée de polysémie réduite correspond à l'hypothèse de monosémie[27] de Ruhl (1989), selon laquelle les mots ont une seule signification de base très abstraite, si l'on fait abstraction des contributions contextuelles

---

[25] Trois sens sont généralement distingués : *bureau* « table de travail », *bureau* « pièce de travail » et *bureau* « lieu de travail » (Fuchs 1996 : 27).

[26] *Faire grève* : « se tenir sur la place de Grève, en attendant de l'ouvrage » (PR). Les ouvriers sans emploi attendaient sur la place de Grève à Paris (au bord de la Seine).

[27] « Monosemic bias : First hypothesis : A word has a single meaning. Second hypothesis : If a word has more than one meaning, its meanings are related by general rules » (Ruhl 1989 : 4).

(linguistiques et extralinguistiques). La polysémie apparente est ramenée à des effets contextuels.

La polysémie « éclatée » (Fuchs 1996), par contre, consiste à réduire la polysémie à l'homonymie et à ramener les cas d'indétermination à des cas d'ambiguïté. Les sens différents sont ramenés à autant de mots différents, la signification est surdéterminée en contexte. Les deux stratégies de polysémie réduite et de polysémie éclatée sont diamétralement opposées (Fuchs 1996), mais elles sont complémentaires (Bianchi 1991). En effet, d'une part, l'approche homonymique (polysémie éclatée) permet à la terminologie traditionnelle de maintenir sa thèse de sens fixe par domaine de spécialité (polysémie éclatée en autant d'homonymes que de domaines de spécialité). D'autre part, à l'intérieur d'un domaine de spécialité, l'approche de polysémie réduite à l'indétermination permet de maintenir le principe de monosémie et d'univocité. L'existence de ces deux stratégies montre aussi que les frontières entre l'homonymie et la polysémie et entre la polysémie et l'indétermination ne sont pas toujours nettes.

Passons finalement aux critères devant permettre de distinguer entre la polysémie et l'indétermination (Tuggy 1993 ; Geeraerts 1993 ; Nerlich et al. 2003). Tuggy (1993) discute d'abord la distinction entre l'indétermination (ou le vague), l'ambiguïté et la polysémie. L'indétermination serait qualifiée d'unité (un sens général et vague), tandis que l'ambiguïté serait caractérisée par la séparation (deux ou plusieurs sens différents). La polysémie se situerait à mi-chemin entre le vague et l'ambiguïté (Tuggy 1993 : 275), dans la mesure où les sens sont à la fois clairement séparés et également reliés, ce qui conduit Tuggy à conclure que la frontière entre le vague et l'ambiguïté est floue. De plus, la distinction entre le vague et la polysémie n'est pas stable, car ce qui semble être un sens différent dans un contexte (polysémie), est réduite à un cas de vague dans un autre contexte (Geeraerts 1993 : 224). En effet, les critères traditionnels (logiques, linguistiques, définitionnels) permettant de distinguer la polysémie (ou l'ambiguïté) et le vague (Cf. 1.2.1.1) ne fonctionnent pas dans ce cas (Geeraerts 1993). Deux par deux, ils mènent à des résultats divergents. Ce qui est polysémie selon un critère, est vague selon un autre critère[28]. Par conséquent, l'ambiguïté et le vague ne doivent pas être considérés comme des catégories classiques avec des frontières nettes et fixes, mais plutôt comme des

---

[28] Par exemple, l'autohyponyme anglais *dog* dans « *Lady is a dog alright, but she is not a dog* » (Geeraerts 1993 : 237), présente une lecture hyperonymique « canis familiaris » et une lecture hyponymique « canis familiaris mâle » (Cf. l'exemple français *homme*). Le test logique permet la négation indépendante des deux lectures, qui indique deux sens polysémiques. Par contre, selon le test définitionnel, la lecture hyponymique relève toujours de la définition (maximalement générale) de la lecture hyperonymique.

catégories prototypiques, avec de meilleurs représentants (membres) et de moins bons représentants. Ainsi, l'appartenance à telle ou telle catégorie n'est pas absolue, mais est une question de gradation (Tuggy 1993 ; Geeraerts 1993), d'où la notion de « polysémie graduée » proposée par Nerlich et al. (2003). Geeraerts (1993) propose dès lors de recourir à un continuum de sens (*continuum of meaning*), plutôt qu'à une dichotomie.

En conclusion, il ressort de ce qui précède que la dichotomie traditionnelle entre la polysémie (caractéristique de la langue générale[29]) et la monosémie (caractéristique de la langue spécialisée) n'est pas opérationnelle. Les critères permettant de distinguer ne fonctionnent pas toujours ou se contredisent même. La littérature révèle également un manque de cohérence au niveau des dénominations et des définitions de ces qualifications sémantiques. Ce qui est ambiguïté (polysémie et homonymie) pour l'un est considéré comme de l'homonymie par l'autre. Quelques études suggèrent de considérer les phénomènes de polysémie comme des patrons sémantiques flexibles et proposent d'élaborer une théorie de polysémie en termes de continuum (Ravin & Leacock 2000) en termes de gradation et de flexibilité (Nerlich et al. 2003).

> We adopt as a working hypothesis the view that almost every word is more or less polysemous, with senses linked to a prototype by a set of relational semantic principles which incorporate a greater or lesser amount of flexibility. (Nerlich et al. 2003 : 8)

D'ailleurs, les phénomènes graduels et progressifs, tels que les changements linguistiques, requièrent des observations statistiques (fréquence d'emploi, degré d'intensité de relations), contrairement aux observations catégorielles (Manning & Schütze 2002). L'idée de gradation sera reprise dans la solution alternative que nous proposons (Cf. 1.2.3).

1.2.2.3 La polysémie dans la langue spécialisée : travaux antérieurs

Récemment, plusieurs études ont démontré qu'il y a de la polysémie dans la langue spécialisée, même à l'intérieur d'un seul domaine spécialisé[30], en s'appuyant principalement sur l'analyse de contextes spécialisés. Citons notamment les travaux

---

[29] Plus de 40% des mots du Petit Robert seraient polysémiques, d'après un calcul statistique approximatif sur le Petit Robert (Fuchs 1996 : 29).

[30] La polysémie nominale et adjectivale apparaît aussi bien dans un corpus de langue spécialisée (par exemple un discours médical entre spécialistes) que dans un corpus de langue générale (par exemple un discours politique à l'adresse du grand public) (Fabre et al. 1997).

de Arnzt et Picht (1989), les travaux de Temmerman (1997, 2000a et 2000b) dans le domaine des sciences de la vie, les expérimentations de Condamines et Rebeyrolle (1997) dans le domaine de l'espace, ainsi que les expérimentations plus récentes de Eriksen (2002) et Ferrari (2002) dans le domaine juridique, respectivement pour l'allemand et l'espagnol.

- *Arntz & Picht (1989)*

Dès 1989, Arntz et Picht ont mis en évidence la présence d'unités polysémiques dans la langue spécialisée (*Fachsprache*), en l'occurrence dans l'allemand technique. Ils observent que les dictionnaires de spécialité (alphabétiques) contiennent des exemples de polysémie[31] et que « le nombre de mots polysémiques augmente proportionnellement avec le nombre de domaines de spécialité traités » (Arntz & Picht 1989 : 135). La polysémie est considérée comme l'attestation de différents sens dans différents sous-domaines de la langue spécialisée technique. Etant donné que la relation entre notion et dénomination n'est pas toujours univoque et encore plus rarement bi-univoque, Arntz et Picht proposent des aides à l'explication ou à la compréhension, notamment le domaine de référence (le sujet), la définition, le contexte (les cooccurrents directs) et l'indication de la source.

- *Temmerman (1997, 2000a et 2000b)*

Comme nous l'avons mentionné ci-dessus, Temmerman (2000a et 2000b) remet en question l'idéal d'univocité, après avoir analysé la sémantique de certains termes dans un corpus anglais du domaine des sciences de la vie. Les données empiriques de son approche sémasiologique permettent de montrer comment le sens des termes peut changer dans le temps. Le terme *cloning*, par exemple, fait l'objet de plusieurs extensions sémantiques en raison de nouvelles inventions. Au fil du temps, le terme *cloning* est surchargé sémantiquement et subit deux types de changements (*shifts*) sémantiques. Premièrement, le terme *cloning* subit un transfert métaphorique du domaine de la biologie au domaine de l'informatique et même à la langue générale : le terme devient un mot (Cf. déterminologisation). A un moment donné, *cloning* revêt donc simultanément le sens général et le sens spécialisé. Deuxièmement, du fait que dans la langue spécialisée, le terme *cloning* est tellement surchargé dans ses emplois plus précis, il est remplacé par le terme *amplification*, un processus d'indexation générique (*generic posting*) (Temmerman 2000a : 149).

---

[31] Par exemple le mot *Lager* signifie : (1) « Wellenlager » (Maschinenbau) ; (2) « Bettung » (Bauwesen) ; (3) « Pfanne » (Werkzeug) ; (4) « Lagerstätte, Fundort » (Bergbau, Geologie) ; (5) « Lagerraum » (Arntz & Picht 1989 : 135).

- *Condamines et Rebeyrolle (1997)*

Dans un corpus de textes spécialisés relevant du domaine de l'espace, Condamines et Rebeyrolle (1997) analysent entre autres le mot *satellite*. Elles insistent d'abord sur la notion de « point de vue » et identifient un premier type de point de vue dans l'actualisation de la langue générale en discours spécialisé, ce qui est un point de vue collectif lié à une connaissance spécifique du domaine. C'est ce type de point de vue qui intéresse particulièrement le terminologue. Le deuxième type de point de vue, individuel, concerne l'actualisation de la langue spécialisée en discours, en fonction d'un locuteur particulier.

Pour étudier la polysémie dans la langue spécialisée, Condamines et Rebeyrolle (1997) ont analysé deux corpus de textes spécialisés du domaine de l'espace, en l'occurrence des documents du Centre National d'Études Spatiales (CNES), de la division « Observation de la terre » et de la division « Mathématiques spatiales ». L'outil ALCESTE d'analyse statistique permet le découpage en plusieurs parties ou classes thématiquement homogènes. Tout comme pour la langue générale, il faudra repérer des termes et classer leurs contextes d'apparition afin d'identifier si ces contextes peuvent être considérés comme sémantiquement homogènes ou non. Il s'agit de « cerner le sens d'un mot en s'appuyant sur les connaissances linguistiques que l'on a sur le contexte » (Condamines & Rebeyrolle 1997 : 178). La « polyacception », c'est « le fait que plusieurs classes sémantiques de contextes puissent être identifiées pour un terme » (ibid.). Cette polyacception est la manifestation de plusieurs points de vue différents.

Pour le mot *satellite*, l'analyse du corpus spécialisé a permis de relever six patrons syntaxico-sémantiques, caractéristiques de six types d'acceptions[32]. Ensuite, un expert du domaine a pu donner, à chacune de ces six acceptions, une identification « que l'on peut considérer comme manifestant des points de vue » (Condamines & Rebeyrolle 1997 : 181). Dans chacun des deux corpus, on a pu identifier un patron syntaxico-sémantique dominant, donc une acception dominante correspondant à un point de vue dominant. Condamines et Rebeyrolle identifient le point de vue dominant comme la manifestation d'un point de vue collectif d'une compétence socioprofessionnelle. Dans le corpus « Observation de la terre », le point de vue

[32] Six acceptions sont attestées : (1) « un corps artificiel » (lancé de la terre de façon à devenir le satellite d'une planète) : le sens du dictionnaire ; (2) « un mobile » (corps qui peut être mu, dont on peut changer la position) : dans ce type d'utilisation, une propriété est privilégiée ; (3) « une plate-forme » : par glissement par métonymie, car la plate-forme est une partie du satellite ; (4) « un véhicule » : par glissement par métonymie ; (5) « un hôte » : une propriété (ou trait dénotatif) est privilégiée ; (6) « un relais », « une interface » : une autre propriété est privilégiée (Condamines & Rebeyrolle 1997 : 181-182).

plate-forme est nettement dominant et dans le corpus « Mathématiques spatiales », c'est le point de vue *mobile* qui est dominant. L'existence de point de vue est d'ailleurs étroitement liée à une connaissance spécialisée particulière. Toutefois, la présence d'un point de vue dominant, par exemple *mobile*, n'empêche pas la présence, dans le discours, de points de vue secondaires (*véhicule, hôte, relais*), correspondant à des points de vue individuels. Condamines et Rebeyrolles mettent ainsi en évidence l'existence de polyacceptions d'un terme polysémique dans la langue spécialisée, par le biais de manifestations linguistiques diverses.

- *Eriksen (2002)*

Eriksen (2002) étudie le mot *Sache* (chose), employé dans la langue générale et dans la langue juridique du droit civil, ainsi que dans différents sous-domaines de l'allemand juridique. Il conclut que, même si la polysémie est plus fréquente dans la langue générale, elle s'observe aussi dans la langue juridique. Le mot *Sache* est défini par la loi allemande comme « ein körperlicher Gegenstand » (un objet physique). En effet, le juriste considère comme *Sachen* des objets solides, liquides et gazeux, tandis que la langue générale ne prend en considération que les objets solides. Deux autres exemples montrent également que le mot *Sache* a un sens différent dans la langue générale et la langue spécialisée. Le mot se caractérise par une catégorisation plus précise et plus fine dans la langue spécialisée, que l'on ne retrouve plus dans la langue courante (Eriksen 2002 : 217-219). Ainsi, les animaux sont considérés comme des choses par les juristes, mais pas dans la langue générale. Depuis 1989, le droit civil considère le logiciel comme un objet physique, tandis que la langue générale le reconnaît rarement comme *Sache*. On observe donc clairement que le sens d'un mot est polysémique selon l'emploi en langue générale ou en langue spécialisée. En plus, le rapport entre la langue générale et la langue spécialisée ne correspond pas toujours à l'opposition entre le vague (l'inexactitude) et la précision (Eriksen 2002 ; von Hahn 1998). Bien que l'expression *un peu plus de 1000 euros* soit moins précise que *plus de 1000 euros* du point de vue linguistique et logique, elle est plus claire et plus précise du point de vue communicatif. Von Hahn (1998) conclut ainsi que dans un texte de vulgarisation, une expression plus exacte du point de vue scientifique peut constituer une entrave à la communication.

Après la première comparaison entre la langue générale et la langue juridique, la deuxième comparaison porte sur plusieurs sous-domaines juridiques et permet de relever ce que l'on qualifie de *Fachpolysémie*, ou, littéralement, de « polysémie de spécialité ». Eriksen (2002) signale d'ailleurs que dans la langue juridique, on ne demande pas d'introduire la normalisation, contrairement aux autres langues spécialisées. Ainsi, le mot *Sache* est employé différemment dans les différents sous-domaines juridiques. Les sens différents ont comme point de départ la définition de

« körperlicher Gegenstand » (objet physique). Ainsi, contrairement au droit civil, le droit administratif ne maintient pas la caractéristique « körperlich ». De même, le logiciel n'est pas une *Sache* dans le droit pénal, même si tel est le cas dans le droit civil. On observe donc une multitude de sens reliés, ayant en commun la même définition de base. Cette multitude de sens ou plurivocité est qualifiée de polysémie, tant entre langue générale et langue spécialisée, en l'occurrence juridique, qu'entre plusieurs sous-domaines de la langue juridique.

- *Ferrari (2002)*

Le phénomène de la polysémie se retrouve, non seulement entre plusieurs sous-domaines d'une langue spécialisée, mais également à l'intérieur d'un sous-domaine de spécialité. Dans le domaine juridique en espagnol, plus particulièrement dans le sous-domaine du droit constitutionnel, Ferrari (2002) observe des phénomènes de variation conceptuelle, que l'on pourrait qualifier de polysémie, malgré le haut degré de précision des textes juridiques. Ferrari étudie les termes espagnols *distinción* et *discriminación* dans un corpus spécialisé de dix traités internationaux, dans le but de fournir des preuves empiriques justifiant la remise en question de l'idée d'univocité et de monosémie des unités terminologiques.

A cet effet, elle donne d'abord les définitions des deux termes dans la langue générale et dans le domaine juridique. Ensuite, elle identifie les contextes syntactico-sémantiques pour vérifier si le signifié des termes est identique dans tous les schémas syntactico-sémantiques ou s'il s'agit de cas de polysémie. Dans le domaine juridique, *distinción* et *discriminación* ont un trait sémantique en commun, à savoir celui de « différenciation ». En plus, *discriminación* fonctionne comme hyponyme de l'hyperonyme *distinción* en raison de son trait sémantique supplémentaire « à des fins de persécution » (Ferrari 2002 : 226). Les deux termes ne sont pas synonymes, car *distinción* s'emploie généralement dans d'autres contextes, marqués par l'absence de compléments indiquant la cause ou les motifs. Dans certains contextes, par contre, les deux termes *distinción* et *discriminación* alternent, ce qui veut dire que *distinción* est employé dans une de ses acceptions ou un de ses sens (Ferrari 2002 : 241). Ainsi, il s'agit clairement d'un cas de polysémie à l'intérieur d'un (sous-)domaine de spécialité. Il a pu être détecté grâce à l'analyse de contextes sémantiques et syntaxiques dans un corpus de textes spécialisés, faite manuellement, pour un nombre limité de termes et pour un nombre limité de contextes (corpus de dix traités internationaux).

Les études et les expérimentations discutées ci-dessus portent généralement sur un seul mot ou terme, ou, tout au plus, sur un nombre limité d'unités lexicales, analysées manuellement. La méthodologie utilisée s'appuie principalement sur l'analyse des contextes linguistiques des unités spécialisées, par le biais de corpus

textuels. Malgré leur champ d'étude limité, ces expérimentations fournissent des indications concrètes sur la présence de polysémie dans la langue spécialisée. Ces résultats convaincants suggèrent que la méthodologie, basée sur l'étude de textes et de contextes spécialisés, mérite d'être appliquée à plus grande échelle afin d'étudier la sémantique, et potentiellement la polysémie, des unités lexicales de textes spécialisés, qu'il s'agisse de termes au sens strict ou non.

1.2.3 Solution alternative : un continuum sémantique

Des deux sections précédentes, il ressort que la dichotomie traditionnelle entre monosémie et polysémie (Cf. 1.2.1) n'est pas opérationnelle (Cf. 1.2.2). L'adage qui veut qu'il n'y a que de la monosémie en langue spécialisée est remis en question par les partisans de la terminologie descriptive. Par ailleurs, il s'est avéré que les critères permettant de distinguer monosémie, polysémie, homonymie et indétermination ne sont ni suffisants ni convergents. En effet, diverses études théoriques sur la polysémie suggèrent l'idée d'un continuum ainsi que l'idée de gradation. Finalement, les études et les expérimentations récentes sur des corpus spécialisés montrent la présence indéniable de la polysémie dans la langue spécialisée.

Par conséquent, nous proposons de situer la monosémie et la polysémie, tout comme l'homonymie et l'indétermination, sur un continuum, qui permettra également de rendre compte des gradations suggérées ci-dessus. Ce continuum pourrait aller de la monosémie (un seul sens nettement délimité), en passant par l'indétermination (un seul sens sous-déterminé, mais précisé par le contexte) et la polysémie (plusieurs sens reliés, désambiguïsés par le contexte) pour finalement en arriver à l'homonymie (plusieurs unités lexicales différentes mais formellement identiques et coïncidantes). Toutefois, le continuum que nous proposons ne consiste pas en une subdivision sémantique classique en 4 parties, avec ou sans gradations, mais sera étudié comme quantification (automatisée) de la sémantique (Cf. chapitre 5).

A la différence des expérimentations récentes citées ci-dessus, observant la polysémie de certaines unités lexicales dans un corpus spécialisé, nous proposons d'étudier ce phénomène à plus grande échelle, pour un vaste ensemble d'unités lexicales d'un corpus spécialisé, en tenant compte des aspects linguistiques pertinents.

1.3 RESTRICTIONS

Avant de passer aux questions et hypothèses de recherche, il convient de préciser un certain nombre de restrictions que nous avons introduites par rapport à l'objet d'étude. Deux questions importantes se posent. La première concerne les unités polylexicales et la deuxième la désambiguïsation automatisée.

La première question découle de l'utilisation du corpus spécialisé et touche à la délimitation de notre objet d'étude. Notre corpus de textes spécialisés contient parmi les unités lexicales spécifiques non seulement des unités lexicales simples (par exemple *machine, usiner, outil, broche*), mais également, et peut-être surtout, des unités lexicales complexes, ou unités polylexicales, qui se composent de plusieurs unités simples, telles que *machine à usiner*. Toutefois, notre étude sémantique quantitative se limite au niveau des unités lexicales simples du corpus spécialisé, bien que la plupart des unités terminologiques[33] se situent à ce niveau d'unités complexes. Plusieurs raisons justifient notre décision.

Nous considérons comme unités simples tous les lemmes (ou formes canoniques) des unités typographiques, tels qu'ils sont identifiés par l'analyseur Cordial (Cf. chapitre 3 sur l'exploitation du corpus). Les unités avec trait d'union ou avec apostrophe sont considérées comme des unités simples, étant des unités typographiques. Cependant, les mots séparés par un espace sont considérés comme deux mots distincts et donc catégorisés par Cordial comme deux lemmes différents. Autrement dit, à ce niveau d'analyse minimale (identification des lemmes), nous adoptons comme critère distinctif l'orthographe, qui est le critère de l'analyseur automatique Cordial : un lemme est ce qui se trouve entre deux espaces[34]. A titre d'exemple, *machine-outil* pourra faire partie des unités spécifiques, tandis que l'unité complexe *machine à usiner* ne sera pas retenue.

Nous nous concentrons dans cette étude sur les unités simples, dans le but de développer une méthodologie opérationnelle et d' appliquer cette méthodologie de façon automatisée. En effet, notre étude vise notamment à développer une mesure permettant de quantifier la monosémie et de calculer le degré de monosémie par le biais d'une formule de recoupement (Cf. chapitre 5). Dans un premier stade, la formule est développée et mise au point pour les unités simples. Nous envisageons, dans un deuxième stade, de l'implémenter également pour les unités polylexicales. Toutefois, cela dépasse les limites que nous nous sommes fixées dans le cadre de la présente étude.

---

[33] Van Campenhoudt (2002b) procède à un dénombrement comparatif de la répartition des termes simples et complexes. Pour le français, les termes complexes constituent 66,7% de toutes les unités terminologiques, tandis que les termes simples n'en représentent que 33,3%. (A noter : terme simple = suite de caractères n'incluant ni espace, ni trait d'union, ni apostrophe).

[34] Pour une justification plus approfondie : voir chapitre 3.

En outre, la restriction de l'objet d'étude aux unités simples s'explique par des raisons de faisabilité informatiques, tant en ce qui concerne le repérage des unités spécifiques (Cf. chapitre 4) que l'identification et l'exploitation de leurs cooccurrences pertinentes (Cf. chapitre 5). Même s'il existe des outils d'extraction terminologique[35], permettant de repérer les unités polylexicales (Bourigault et al. 2001), ces unités complexes posent problème lors du calcul des spécificités. Pour l'instant, il n'est guère possible de déterminer le degré de spécificité des unités complexes de façon fiable et statistiquement significative (Cf. chapitre 4). D'ailleurs, il convient de s'interroger sur la pertinence des techniques d'extraction automatique de termes pour notre étude. Ces techniques s'appuient généralement sur un algorithme hybride avec une composante syntaxique importante, c'est-à-dire des structures syntaxiques récurrentes (Lemay, L'Homme & Drouin 2005)[36]. Ainsi, plusieurs variables concourent au repérage des unités terminologiques complexes plutôt qu'une seule. Cependant, notre recherche, et plus particulièrement l'analyse de régression à laquelle nous procédons, requièrent une seule variable linguistique, c'est-à-dire un critère de spécificité clair et précis. Par conséquent, il est plus prudent, à titre provisoire, de restreindre l'analyse aux unités lexicales simples.

La deuxième restriction, qui s'applique à la désambiguïsation, découle de l'analyse sémantique. En effet, la plupart des analyses sémantiques automatisées ont pour objet la désambiguïsation automatique ou la WSD (*Word Sense Disambiguation*). Elles s'inscrivent dans des projets d'évaluation des résultats de techniques de désambiguïsation, tels que Senseval et Romanseval (Cf. chapitre 5). Or, le but de notre étude sémantique n'est pas de déterminer le nombre de sens des unités polysémiques, ni d'identifier les différents sens en question. Nous proposons en revanche d'adopter une approche de la polysémie, qui permettra d'élaborer une mesure pour calculer le degré de monosémie. Il s'agit donc principalement de quantifier l'analyse sémantique. Ainsi, nous espérons contribuer à l'analyse sémantique automatique, domaine qui est en pleine évolution.

---

[35] Signalons notamment l'outil INTEX – UNITEX (http://www-igm.univ-mlv.fr/~unitex/) et l'outil LEXTER (Bourigault 1994 ; Bourigault et al. 2001).

[36] Les techniques d'extraction automatique d'unités polylexicales reposent généralement sur la combinaison de stratégies linguistiques (patrons syntaxiques récurrents) et de stratégies statistiques (calculs statistiques) (Lemay, L'Homme & Drouin 2005).

Chapitre 2
Questions et hypothèses de recherche

Le deuxième chapitre a pour but d'expliciter et de justifier les questions auxquelles nous tenterons de répondre dans les chapitres suivants ainsi que de formuler des hypothèses. Dans une première partie, nous préciserons les objectifs de recherche et la justification méthodologique (2.1). La deuxième partie sera consacrée à la question principale (2.2) : la corrélation entre le continuum de spécificité et le continuum de monosémie. Comme d'autres facteurs influent également sur la monosémie, ces facteurs feront l'objet de questions complémentaires, explicitées dans la troisième partie (2.3). Finalement, la dernière partie de ce chapitre portera sur des questions détaillées (2.4), c'est-à-dire les classes lexicales et les sous-corpus.

2.1 OBJECTIFS DE RECHERCHE ET JUSTIFICATION

Comme nous l'avons précisé plus haut, ce travail est une étude sémantique quantitative d'un corpus spécialisé. Les trois adjectifs (*sémantique*, *quantitative* et *spécialisé*) méritent un mot d'explication, puisque notre étude vise à remettre en question la thèse *sémantique* du monosémisme préconisée par l'approche traditionnelle au moyen d'une étude *quantitative* d'un corpus *spécialisé* (2.1.1). Notre étude s'appuie principalement sur le degré de spécificité et sur le degré de monosémie (2.1.2). A cet effet nous développerons une mesure du degré de monosémie, permettant de quantifier et d'objectiver l'analyse sémantique du corpus spécialisé. Dans la dernière section (2.1.3), nous mettrons en lumière l'originalité de notre étude.

2.1.1 Remise en question de la thèse monosémiste : étude quantitative

Nous commencerons par expliciter les trois adjectifs qualificatifs mentionnés ci-dessus, dans le but d'expliquer et de justifier les objectifs de recherche. L'étude sera conduite sur un corpus *spécialisé*, en l'occurrence un corpus de textes relevant du domaine technique des machines-outils pour l'usinage des métaux. Une étude linguistique, qui se focalise sur un domaine technique, soulève tout de suite des questions sur les particularités de la langue spécialisée utilisée dans le domaine en

question. Dans la langue spécialisée, les besoins communicatifs des spécialistes requièrent plus de précision, ce que la terminologie traditionnelle définit comme l'univocité, la monoréférentialité et la monosémie des unités terminologiques de la langue spécialisée (Cf. 1.2.1.3).

Cette caractéristique traditionnelle de la monosémie des unités terminologiques d'un corpus spécialisé justifie le deuxième adjectif de notre étude, à savoir *sémantique*. Généralement, une étude sémantique s'interroge sur le sens. Les phénomènes de monosémie, de polysémie, d'homonymie ou d'indétermination des unités lexicales (ou grammaticales) y ressortissent. L'objectif principal de notre étude sémantique est de vérifier si les unités lexicales de notre corpus technique sont monosémiques, comme le prétendent les monosémistes traditionnels ou, par contre, s'il existe des unités lexicales polysémiques, comme le suggèrent les partisans de la terminologie descriptive. Pour évaluer la thèse monosémiste de l'approche traditionnelle, en ayant recours à la linguistique de corpus, il faudra opérationnaliser la thèse monosémiste et la reformuler en une question opérationnelle et mesurable, ce qui permet de justifier le troisième et dernier aspect de notre étude, à savoir la dimension *quantitative*. S'il est vrai que les unités lexicales de la langue spécialisée (d'un corpus technique) sont monosémiques, ce sera d'autant plus vrai pour les unités lexicales les plus spécifiques et les plus représentatives de ce corpus technique. Par conséquent, nous nous demandons si les unités lexicales[37] les plus spécifiques du corpus technique sont effectivement les plus monosémiques.

L'idée que les unités lexicales, qui sont plus ou moins spécifiques, sont plus ou moins monosémiques, implique l'idée de gradation ou de continuum, suggérée dans le chapitre précédent. Nous proposons dès lors d'opposer au classement catégoriel des unités lexicales (termes versus mots) un continuum de spécificité. Celui-ci comprend seulement des unités lexicales spécifiques du corpus technique, allant des unités lexicales les plus spécifiques aux moins spécifiques. Le classement catégoriel sur le plan sémantique (monosémie versus polysémie) est remplacé par un continuum sémantique, allant des unités les plus monosémiques aux unités les moins monosémiques ou, ce qui revient au même, les plus polysémiques. Afin d'évaluer la thèse des monosémistes traditionnels, nous procédons donc à des évaluations fondées sur cette double dimension, impliquant des gradations en termes de degré de spécificité et de degré de monosémie.

[37] Il est à noter que les unités grammaticales seront supprimées de la liste des spécificités, qui ne comprendra que des unités lexicales (Cf. chapitre 4).

2.1.2 Le degré de spécificité et le degré de monosémie

Les deux grands axes méthodologiques seront, d'une part, l'axe de l'identification des spécificités (Cf. chapitre 4), et d'autre part, l'axe de la quantification de la monosémie (Cf. chapitre 5). L'axe des spécificités permettra d'attribuer un degré de spécificité, qui indiquera à quel point les unités lexicales du corpus technique sont spécifiques. L'axe sémantique attribuera un degré de monosémie, qui indiquera à quel point les spécificités sont monosémiques.

Pour identifier les unités les plus spécifiques du corpus technique, c'est-à-dire les « spécificités » ou les « mots-clés », nous allons confronter le corpus technique à un corpus de référence de langue générale, à l'aide de la méthode des mots-clés (Cf. chapitre 4 pour les détails techniques et méthodologiques). En effet, les spécificités ne sont pas simplement les unités linguistiques les plus fréquentes du corpus technique, mais les unités linguistiques les plus caractéristiques et les plus représentatives du corpus technique. En termes relatifs, les spécificités sont significativement plus fréquentes dans le corpus technique que dans un corpus de référence de langue générale.

A titre d'exemple, nous visualisons la comparaison simplifiée d'un corpus spécialisé (200 mots) et d'un corpus de référence plus étendu (500 mots) (Cf. figure 2.1). Les lettres représentent des mots. Le mot *a*, indiqué en rouge, est significativement plus fréquent dans le corpus spécialisé (9 fois) que dans le corpus de référence (1 fois), compte tenu de la taille des deux corpus. Le mot *e* est aussi fréquent dans le corpus spécialisé (9 fois), mais il est aussi très fréquent dans le corpus de référence (15 fois). Dès lors, le mot *e* ne sera pas spécifique du corpus spécialisé, puisque sa fréquence relative dans le corpus spécialisé (9/200) est comparable à sa fréquence relative dans le corpus de référence (15/500). Le mot *t* ne sera pas non plus spécifique du corpus spécialisé, puisque sa fréquence relative dans le corpus de référence (12/500) est supérieure à sa fréquence relative dans le corpus spécialisé (2/200).

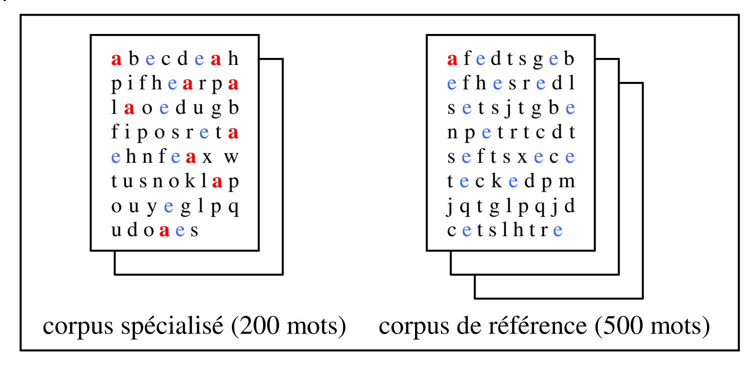


Figure 2.1 Visualisation des spécificités d'un corpus spécialisé

De même, pour identifier les spécificités du corpus technique, nous comparons la fréquence relative d'une unité linguistique dans le corpus technique à sa fréquence relative dans le corpus de référence de langue générale. Cette comparaison permet aussi de déterminer le degré de spécificité de cette unité linguistique (Cf. chapitre 4), car plus l'unité linguistique est spécifique du corpus technique par rapport au corpus de référence de langue générale, plus son degré de spécificité sera élevé. Le degré de spécificité permettra en outre d'ordonner les spécificités et de les situer sur une échelle (ou un continuum) de spécificité. Notons d'emblée que les unités les plus spécifiques sont généralement très fréquentes[38] (par exemple *machine, outil, usinage, pièce*, etc.).

Pour déterminer le degré de monosémie des spécificités du corpus technique, nous procéderons à l'analyse des cooccurrences (Cf. chapitre 5). Celle-ci permettra de quantifier la monosémie en implémentant la monosémie en termes d'homogénéité sémantique. En effet, une unité lexicale monosémique apparaît dans des contextes plutôt homogènes sémantiquement, c'est-à-dire qu'elle se caractérise par des cooccurrents qui appartiennent à des champs sémantiques similaires. Par contre, une unité lexicale polysémique se caractérise par des cooccurrents plus hétérogènes sémantiquement, appartenant à des champs sémantiques différents. L'accès à la sémantique des cooccurrents d'un mot de base se fait à partir de leurs cooccurrents, c'est-à-dire à partir des cooccurrents de deuxième ordre. Si les cooccurrents d'un mot de base partagent beaucoup de cooccurrents de deuxième ordre, ces derniers se recoupent formellement, ce qui est une indication de l'homogénéité sémantique des cooccurrents du mot de base. Ainsi, le degré de ressemblance ou de similarité lexicale des cooccurrents d'un mot de base est proportionnel au degré de monosémie de ce mot de base. Par conséquent, un recoupement important des cooccurrents de deuxième ordre révèle un degré plus important de monosémie du mot de base.

En guise d'exemple, nous visualisons l'unité lexicale spécifique *tour*, indiquée en gras dans deux contextes différents (Cf. figure 2.2 : phrases (1) et (2)). Il est évident que *tour* n'est pas une unité lexicale monosémique : on constate qu'elle se caractérise par des cooccurrents hétérogènes sémantiquement. En effet, les cooccurrents *usine* et *minute* (indiqués en gras et soulignés) relèvent d'un champ sémantique différent. Ils indiquent les deux sens différents de l'unité lexicale *tour*, à savoir « machine-outil pour l'usinage des pièces » (cooccurrent (1) : *usine*) et

---

[38] Par contre, les unités les plus fréquentes du corpus technique ne sont pas nécessairement des unités spécifiques. Ainsi, les unités grammaticales *de, le, à, pour,* etc. sont très fréquentes dans le corpus technique, mais elles sont également très fréquentes dans le corpus de référence de langue générale. Par conséquent, ces unités ne sont pas significativement plus fréquentes dans le corpus technique et elles ne sont pas des unités spécifiques.

« rotation, révolution » (cooccurrent (2) : *minute*). Pour avoir accès à la sémantique du cooccurrent *usine*, par exemple, on analysera ses cooccurrents (soulignés), non seulement dans cette phrase (1) (*alésage*, *centre*, etc.), mais également dans les autres contextes d'apparition d'*usine*, par exemple dans la phrase (3) (*outils, pièces*, etc.). L'analyse porte donc sur tous les cooccurrents pertinents (donc sur tous les cooccurrents pertinents de deuxième ordre) de tous les cooccurrents pertinents (*usine, minute*, etc.) d'un mot de base (*tour*). Cette analyse permettra de vérifier à quel point les cooccurrents des cooccurrents sont partagés, c'est-à-dire dans quelle mesure ils se recoupent. Ainsi, leur degré de recoupement sera une indication du degré de monosémie du mot de base.

(1) La première est un **tour** sur lequel on **usine** l'alésage central. Ensuite, un centre d'usinage usine la forme de l'une des extrémités qui ressemble à une fleur à huit pétales.

(2) …des broches pouvant monter jusqu'à quinze mille **tours** par **minute**, voire plus, puisque cette technologie ...

…

(3) Un tour CNC équipé d'outils modulaires Capto **usine** les pièces en question avec une vitesse de coupe de 150 m/mn…

(4) La pièce tourne sur le tour à une certaine vitesse de broche (n), exprimée en tours par **minute** (tr/mn).

Figure 2.2 Visualisation des cooccurrents d'une unité lexicale spécifique

Nous déterminerons donc le degré de monosémie des spécificités à partir du degré de recoupement des cooccurrents de leurs cooccurrents, que nous calculons à partir d'une mesure de recoupement (Cf. chapitre 5 pour les détails techniques et méthodologiques). Une fois obtenu, le degré de monosémie permettra de situer les spécificités sur une échelle d'homogénéité sémantique (ou de monosémie).

Afin d'évaluer la thèse monosémiste de l'approche traditionnelle, nous proposons de la reformuler en une question opérationnelle et mesurable, conduisant à une analyse quantitative et statistique. La question se pose donc de savoir s'il y a une corrélation entre, d'une part, le continuum de spécificité et, d'autre part, le continuum de monosémie. Notons d'emblée que des recherches supplémentaires s'imposent pour examiner la relation précise entre, d'une part, notre mesure de monosémie, implémentant la monosémie comme homogénéité sémantique, et, d'autre part, ce que l'on considère traditionnellement comme monosémie ou polysémie. Nous recourons à cette mesure de monosémie ou mesure de recoupement, dans le but de

développer un critère opérationnalisable et mesurable. Sans recherche supplémentaire, il serait impossible d'affirmer que notre mesure de monosémie et les degrés de monosémie calculés correspondent parfaitement à ce que les terminologues traditionnels considèrent comme monosémie ou polysémie.

2.1.3 Originalité du travail

L'originalité de cette étude réside principalement dans le développement d'une mesure permettant d'évaluer le degré de monosémie. Cette mesure permettra, non seulement de quantifier la monosémie et d'automatiser l'analyse sémantique, mais également de procéder à des analyses statistiques en vue de fournir des réponses objectives aux questions posées par la présente recherche. De par son approche, notre étude vise à réconcilier la linguistique et la technique (notamment l'informatique et la statistique). Elle recourt à la technique pour mieux comprendre et expliquer certains aspects de la linguistique, comme nous verrons dans les chapitres suivants. Notre étude se situe donc au carrefour de trois disciplines : la linguistique de corpus, l'informatique et la statistique.

En plus, l'approche quantitative et automatisée adoptée sera mise à l'épreuve à grande échelle, étant donné que l'analyse empirique porte sur presque 5000 mots d'un corpus technique, contrairement aux travaux antérieurs (Condamines & Rebeyrolle 1997 ; Temmerman 2000a ; Eriksen 2002 ; Ferrari 2002). Ces travaux étudient, comme nous, la polysémie dans un corpus représentatif d'un domaine spécialisé, mais ils se limitent à quelques mots seulement.

2.2 QUESTION PRINCIPALE

> Y a-t-il une corrélation entre, d'une part, le continuum de spécificité et, d'autre part, le continuum de monosémie ?

Cette question constitue le point de départ de notre analyse, qui étudiera environ 5000 spécificités d'un corpus technique. En réponse à cette question, nous avançons l'hypothèse que, contrairement à la thèse traditionnelle, les mots (les plus) spécifiques du corpus technique ne sont pas nécessairement (les plus) monosémiques. En effet, certaines unités lexicales spécifiques du corpus technique sont des mots à sens multiples. Citons par exemple le mot *broche* (1) « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » et (2) « outil servant à usiner des pièces métalliques ». Signalons également le mot *découpe* (1) « action de découper » et (2) « résultat de la découpe (pièce découpée) », où les deux sens indiquent clairement une relation métonymique.

Pour étudier la question centrale de notre travail (corrélation entre le rang de spécificité et le rang de monosémie), nous recourons à une analyse statistique de régression simple. Cette analyse de régression simple fera intervenir le rang de monosémie comme variable dépendante (ou variable expliquée) et le rang de spécificité comme variable indépendante (ou variable explicative). Le but est d'expliquer la variation du rang de monosémie en fonction de la variation du rang de spécificité. Autrement dit, le but est de vérifier si le rang de spécificité permet de prédire le rang de monosémie, donc s'il y a une corrélation significative entre les deux variables. Si la thèse monosémiste se vérifie (à savoir la monosémie des unités lexicales de la langue spécialisée), il y aura une corrélation positive entre le rang de spécificité et le rang de monosémie, les mots les plus spécifiques du corpus technique étant les plus monosémiques. Si, par contre, il n'y a pas de corrélation ou si la corrélation est négative, la thèse des monosémistes se trouvera infirmée. Nous soutenons l'hypothèse qu'il n'y a pas de corrélation positive entre le rang de spécificité et le rang de monosémie. Nous avançons donc que les mots les plus spécifiques ne sont pas les plus monosémiques, en remettant en question la thèse monosémiste traditionnelle.

Comme nous l'avons mentionné ci-dessus (Cf. 2.1.2), le rang de monosémie est attribué en fonction du degré de monosémie, qui s'appuie sur le degré de recoupement des cooccurrents de deuxième ordre (les cooccurrents des cooccurrents). Toutefois, il est intéressant, lors du calcul de recoupement, de tenir compte également de la spécificité ou technicité de ces cooccurrents de deuxième ordre. Un facteur de pondération permettra d'inclure la technicité des cooccurrents de deuxième ordre et d'élaborer une mesure de monosémie technique pondérée. Ainsi, la mesure de monosémie, déterminant le degré et donc le rang de monosémie des unités lexicales spécifiques, sera complétée et précisée par une mesure de monosémie technique. Par conséquent, l'analyse principale sera complétée par une analyse de régression simple supplémentaire. Cette analyse fera intervenir le rang de monosémie technique comme variable dépendante (ou expliquée) et elle maintient le rang de spécificité comme variable indépendante (ou explicative). Il est clair que la deuxième analyse de régression simple conduira à nuancer les résultats de l'analyse de régression de base. Ainsi, la question se pose de savoir si la nouvelle mesure de recoupement a plus d'impact sur le degré de monosémie des unités les plus spécifiques ou si, par contre, elle a plus d'impact sur le degré de monosémie des unités moins spécifiques.

Signalons que nous cherchons également à préciser et à nuancer le niveau de la spécificité. A cet effet nous essaierons de développer une variable supplémentaire qui permette de déterminer la technicité d'une unité lexicale, à partir de la différence ou de l'écart entre sa fréquence dans le corpus technique et sa fréquence dans le corpus de référence de langue générale.

2.3 QUESTIONS COMPLÉMENTAIRES

Le rang de monosémie d'un mot n'est pas uniquement influencé par le rang de spécificité, mais également par d'autres facteurs comme sa fréquence dans le corpus technique, sa fréquence dans un corpus de référence de langue générale, sa longueur, sa classe lexicale et le nombre de classes lexicales auxquelles appartient le mot. Dès lors, il est intéressant d'étudier l'impact de ces différents facteurs dans autant d'analyses de régression simple. Dans ces analyses, la variable dépendante sera le rang de monosémie de l'unité lexicale spécifique et la variable indépendante sera un des facteurs cités ci-dessus.

Cependant, ces facteurs peuvent interagir : deux ou plusieurs facteurs peuvent se renforcer ou s'affaiblir, ils peuvent être colinéaires et donc expliquer (en partie) la même variation du rang de monosémie. Par conséquent, il est nécessaire de faire intervenir simultanément tous les facteurs pouvant influer sur le rang de monosémie. Une analyse de régression multiple permettra d'intégrer tous les facteurs pertinents (c'est-à-dire toutes les variables indépendantes) et d'évaluer leur impact combiné sur le rang de monosémie. La question principale sera ainsi complétée par l'étude de questions complémentaires qui font intervenir plusieurs facteurs susceptibles d'influer sur le rang de monosémie. Il est à noter que ces analyses de régression seront conduites, tant pour le rang de monosémie que pour le rang de monosémie technique. Voici donc les questions complémentaires de notre recherche :

> Y a-t-il une corrélation entre, d'une part, chacun des autres facteurs pertinents et, d'autre part, le rang de monosémie ? Quel est l'effet combiné de tous les facteurs sur le rang de monosémie ? Quel facteur rend le mieux compte de la variation du rang de monosémie ?

2.4 ANALYSES DÉTAILLÉES

Nous nous proposons également d'étudier des sous-ensembles de la liste de 5000 spécificités et donc d'y effectuer des analyses de régression simple et multiple détaillées. A cet effet, les spécificités seront réparties en plusieurs sous-ensembles, c'est-à-dire par classe lexicale (substantifs / adjectifs / verbes / adverbes).

> Y a-t-il une corrélation entre le continuum de spécificité et le continuum de monosémie, pour les substantifs, les adjectifs, les verbes et les adverbes ? Quel est l'effet combiné de tous les facteurs pertinents sur le rang de monosémie pour les substantifs, les adjectifs, les verbes et les adverbes ?

Ces analyses détaillées permettront de vérifier la corrélation entre le rang de spécificité et le rang de monosémie en fonction des classes lexicales. Premièrement, elles situeront le rang de spécificité et le rang de monosémie des spécificités d'une classe lexicale déterminée par rapport à l'ensemble des 5000 spécificités. Deuxièmement, elles détermineront les corrélations des nouveaux rangs de spécificité et de monosémie à l'intérieur de la classe lexicale qui fait l'objet de l'analyse détaillée. Selon la théorie traditionnelle, on serait amené à croire, par exemple, que les verbes sont peu spécifiques dans un corpus technique et qu'ils sont donc moins monosémiques que les substantifs. En effet, les études théoriques sur la langue spécialisée affirment que les textes techniques se caractérisent notamment par une surabondance de substantifs, de substantifs déverbaux et d'abréviations et sigles. Nous vérifions à l'aide des analyses de régression détaillées, par classe lexicale et par sous-catégorie, si ces affirmations se confirment dans notre corpus technique. L'analyse plus détaillée des unités lexicales spécifiques avec trait d'union (-) et avec barre oblique (/) constitue un premier pas dans la direction de l'étude des unités polylexicales.

Il est également intéressant de procéder à des analyses de régression simple et multiple pour les différents sous-corpus (revues électroniques / fiches techniques / normes et directives / manuels) et de poser les questions suivantes :

> Y a-t-il une corrélation entre le continuum de spécificité et le continuum de monosémie pour les différents sous-corpus ? Quel est l'effet combiné de tous les facteurs pertinents sur le rang de monosémie pour les différents sous-corpus ?

Pour évaluer la thèse monosémiste de l'approche traditionnelle prescriptive et normative, le sous-corpus des normes et directives est un corpus particulièrement intéressant, parce que ce genre de textes sont censés être prescriptifs et normatifs. Il s'agira de vérifier si ce sous-corpus se distingue des autres sous-corpus, dans ce sens qu'il y aurait une meilleure corrélation ou peut-être une corrélation positive entre le rang de spécificité et le rang de monosémie.

Afin d'approfondir et de nuancer les résultats des analyses de régression pour les normes et directives, ce sous-corpus sera également comparé aux trois autres sous-corpus (revues électroniques / fiches techniques / manuels), qui feront fonction de corpus de référence, tant pour le calcul des spécificités que pour le calcul de la mesure de monosémie technique.

PARTIE II

Corpus et méthodologie

Chapitre 3
Corpus technique et corpus de référence

Le troisième chapitre constitue l'introduction à la partie méthodologique et décrit le corpus technique et le corpus de référence. Les deux axes méthodologiques qui seront expliqués dans les deux chapitres suivants (chapitres 4 et 5), à savoir le calcul du degré de spécificité et celui du degré de monosémie (technique), s'appuient tous les deux sur le corpus technique ainsi que sur le corpus de référence. Pourquoi faut-il recourir à deux corpus ? Le corpus technique constitue le corpus d'analyse ou le corpus de base, sur lequel les analyses sémantiques seront conduites. Toutefois, pour pouvoir déterminer les spécificités, c'est-à-dire les unités lexicales les plus représentatives du corpus technique de langue spécialisée, il faut comparer le corpus technique à un corpus de langue générale. Par conséquent, le corpus de référence de langue générale s'avère indispensable.

Dans ce chapitre, nous expliciterons la constitution et l'exploitation du corpus technique et du corpus de référence. La constitution (3.1) sera expliquée en fonction des principaux critères de constitution de corpus spécialisés. Pour la description de l'exploitation (3.2), nous nous limiterons aux points essentiels, les détails techniques étant joints en annexe. La dernière partie décrira la préparation aux analyses (3.3), notamment la génération des listes de fréquence.

3.1 CONSTITUTION

3.1.1 Constitution du corpus technique

Notre étude s'appuie principalement sur un corpus technique constitué de textes techniques authentiques, que nous avons recueillis nous-même. La constitution de ce corpus n'a pas été une tâche aisée, en raison de la délimitation préalable du domaine spécialisé et des sujets à prendre en considération. Pour la langue spécialisée, il existe très peu de corpus préconstitués facilement accessibles et disponibles, du moins pour le français technique.

Nous expliquons donc les caractéristiques de notre corpus technique à partir des critères de constitution de corpus spécialisés[39], tels qu'ils ont été définis dans Pearson (1998 : 58-62) et dans Bowker & Pearson (2002 : 45-52) (Cf. les sections 3.1.1.1 à 3.1.1.5). Un corpus est une « collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage » (Sinclair 1996, repris dans Habert et al. 1997 : 11). Après avoir analysé de nombreuses définitions de corpus, Pearson (1998) relève les notions-clés de *collection*, *échantillon* et *représentativité*. Bowker et Pearson définissent un corpus comme « a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria » (Bowker & Pearson 2002 : 9). Ainsi, quatre caractéristiques se dégagent de cette définition, à savoir *authentique, électronique, large* et *critères spécifiques*. Les notions-clés et les caractéristiques correspondent aux principaux critères pour la constitution d'un corpus de langue spécialisée. Nous proposons de regrouper les critères de Pearson (1998), étant donné qu'ils se recoupent parfois, et de les compléter, si besoin en est (Bowker & Pearson 2002 ; Habert et al. 1997).

3.1.1.1 La taille et la représentativité

Notre corpus technique est une large collection de textes spécialisés authentiques et électroniques, relevant du domaine des machines-outils pour l'usinage des métaux. Il comprend 1.751.800 occurrences. Selon Pearson (1998), la taille idéale d'un corpus spécialisé serait d'environ 1 million d'occurrences. Bowker & Pearson (2002) mentionnent que la taille des corpus spécialisés bien constitués varie entre une dizaine de milliers et plusieurs centaines de milliers d'occurrences. A titre d'exemple, Jacques (2003)[40] et Valente (2002)[41] ont récemment mené des études sur des corpus spécialisés de 85.000 et de 202.000 occurrences (Jacques 2003) et de 500.000 occurrences (Valente 2002). Notons également que Pearson (1998)[42] a constitué trois corpus spécialisés, respectivement de 4,7 millions de mots (experts-initiés), de 1 million de mots (professeurs-étudiants) et de 230.000 mots (entre experts). Les différences de taille n'avaient pas d'effet significatif sur les résultats, surtout influencés par le contexte communicatif et par la technicité (Pearson 1998 : 64-65).

[39] A l'instar de Valente (2002) et Van Campenhoudt (2002b).

[40] Etude sur la réduction des termes complexes dans les textes spécialisés (Jacques 2003).

[41] Etude sur la remodulation du sens dans un discours spécialisé (Valente 2002).

[42] Etude sur les informations exprimées dans les définitions (Pearson 1998).

Bien entendu, la taille du corpus dépend du domaine spécialisé et du sujet[43], de la disponibilité du matériel sous forme électronique et des objectifs de recherche. A l'ère de la disponibilité de documents électroniques (spécialisés) sur Internet, la constitution de corpus (spécialisés) se trouve certainement facilitée. Toutefois, la prudence s'impose, car il faut veiller à la qualité des textes et des sites (Cf. 3.1.1.2 des textes écrits). Les objectifs de recherche déterminent également la taille du corpus, pour que l'on puisse éviter des problèmes de rareté des données. Ainsi, l'étude de phénomènes lexicaux requiert un corpus plus étendu que l'étude de patrons syntaxiques fréquents. Il faudra donc veiller à assurer la répétition des termes importants du domaine (Cf. Huot 1996).

La taille du corpus soulève également la question de la représentativité du corpus. Le corpus spécialisé est censé refléter la réalité langagière dans le domaine spécialisé. Toutefois, la question de savoir comment on détermine la taille d'un échantillon représentatif reste toujours sans réponse (Pearson 1998). Pour garantir la représentativité de notre corpus technique et la couverture du domaine spécialisé, nous avons recueilli des textes de 11 sources différentes, datant de 1996 à 2002. Le corpus technique se compose de quatre sous-corpus (Cf. figure 3.1), constitués chacun de deux, trois ou quatre sources différentes (Cf. figure 3.2 et tableau 3.1).

- Revues techniques électroniques 790.680 occurrences
- Fiches techniques 296.650 occurrences
- Normes ISO et directives 286.139 occurrences
- Guides et manuels numérisés 378.331 occurrences

La constitution du corpus a été soumise à un expert du domaine[44] afin de juger la pertinence et la représentativité du corpus pour le domaine des machines-outils pour l'usinage des métaux. Notons que le sous-corpus des revues électroniques est plus étendu que les trois autres sous-corpus, qui sont de taille comparable. Cela s'explique principalement par des raisons d'accessibilité et de disponibilité du matériel. Les revues techniques spécialisées, les fiches techniques, ainsi que les directives ont été trouvées sur Internet, sur des sites professionnels et spécialisés.

---

[43] « Plus le sujet est pointu, plus la sélection est restrictive, plus la taille du corpus sera limitée » (Van Campenhoudt 2002b : 4).

[44] Le Prof.dr.ir. J.-P. Kruth (K.U.Leuven), Faculté des Sciences de l'Ingénieur, Département de Mécanique, Division PMA (processus de production), bilingue néerlandais – français.

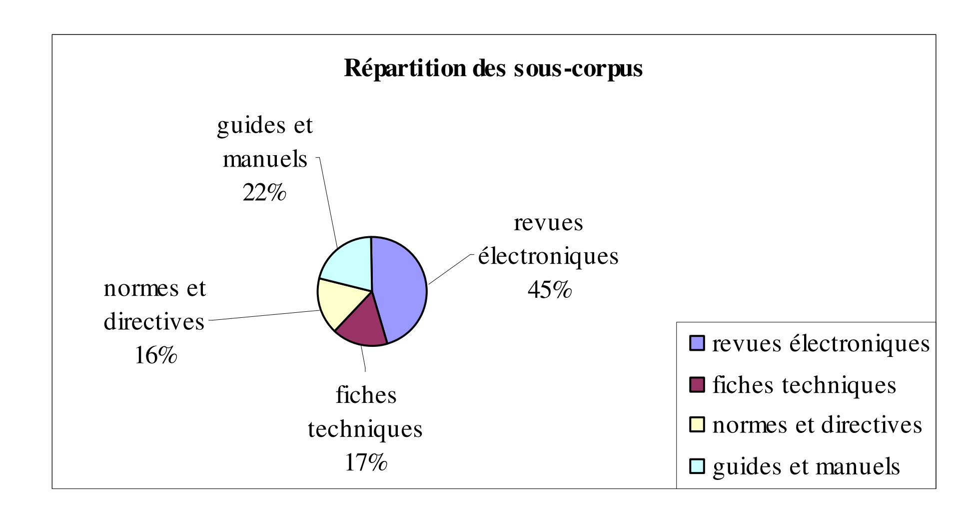


Figure 3.1 Constitution du corpus technique : répartition des sous-corpus

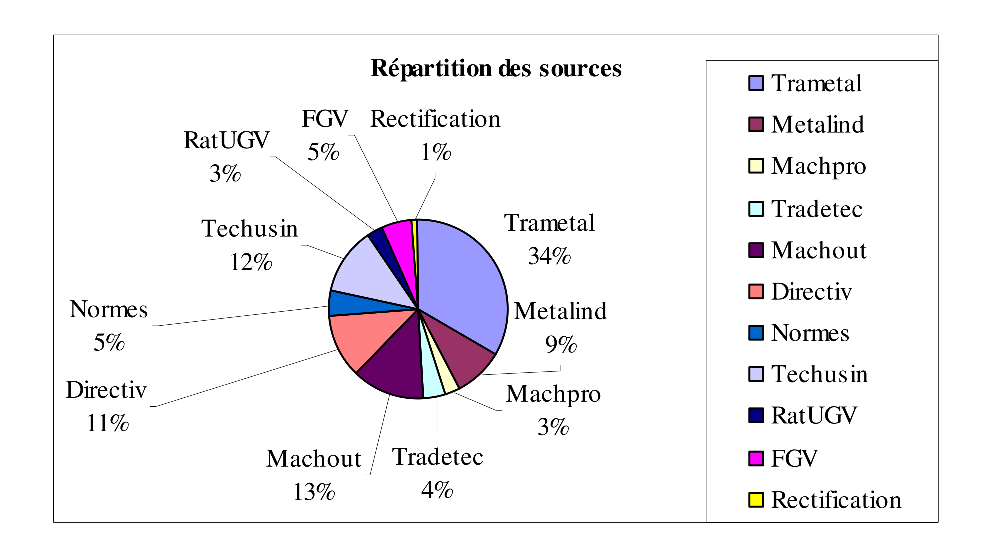


Figure 3.2 Constitution du corpus technique : répartition des sources

La constitution détaillée en fonction des sujets traités dans les différentes sources sera expliquée ci-dessous (Cf. 3.1.1.4 le sujet).

Revues électroniques	
Trametal	Trametal
Metalind	Métal Industries
Machpro	Machines production
Fiches techniques	
Tradetec	Trametal détectés
Machout	Machines-outils
Normes et directives	
Normes	Normes ISO européennes
Directiv	Directives Machines
Guides et manuels	
Techusin	Techniques modernes d'usinage
RatUGV	Rationalisation de l'usinage très grande vitesse
FGV	Fraisage à grande vitesse
Rectification	Rectification des pièces de révolution

Tableau 3.1 Constitution du corpus technique : 11 sources

Les textes des quatre sous-corpus se situent à différents niveaux de normalisation et de vulgarisation, ce qui assure la représentativité et la qualité du corpus[45]. Les deux sous-corpus des normes et directives et des guides et manuels sont plus normatifs et prescriptifs que les deux autres sous-corpus issus des revues électroniques et des fiches techniques, qui sont plus descriptifs. Les normes et directives, les fiches et aussi les revues s'adressent plutôt à des professionnels, tandis que les guides et manuels (et dans une certaine mesure les revues) sont plus didactiques et vulgarisants et visent un public d'étudiants et de semi-experts (Cf. 3.1.1.3 le public-cible). Comme cette étude procède à une remise en question de la thèse monosémiste de l'approche traditionnelle normative et prescriptive, il est indispensable d'inclure dans le corpus technique des normes et des textes normatifs et prescriptifs. En effet, la question est de savoir si les résultats trouvés dans ce sous-corpus permettront de corroborer les résultats des analyses dans le corpus technique entier et dans les autres sous-corpus.

La notion d'échantillon mérite aussi qu'on s'y attarde un moment pour ce qui est de la représentativité du corpus. Un corpus spécialisé est considéré comme un échantillon représentatif de la langue du domaine spécialisé. Une étude conduite sur l'échantillon devrait donc permettre des généralisations et des extrapolations. Toutefois, à ce sujet, la prudence s'impose également, étant donné que les généralisations ne sont pas toujours fiables. Lorsqu'on étudie par exemple un corpus d'un domaine appartenant aux sciences pures, on ne peut pas généraliser et

[45] « The results are only as good as the corpus » (Sinclair 1991 : 13).

extrapoler les résultats à un autre domaine spécialisé, car la langue des sciences et techniques, et a fortiori la langue des sciences pures, a ses particularités (Jacques 2003). Par conséquent, nous n'envisageons pas de généraliser les résultats de cette étude sémantique à tous les domaines techniques, ni à tous les domaines spécialisés. Les résultats de l'analyse calculés sur notre corpus technique seront valables uniquement pour le corpus des machines-outils pour l'usinage des métaux. Bien évidemment, la méthodologie est parfaitement transposable à d'autres corpus spécialisés et à d'autres domaines.

3.1.1.2 Des textes écrits

En ce qui concerne le code des textes (oral ou écrit), Pearson (1998) recourt uniquement à des textes spécialisés écrits, qu'elle inclut dans leur totalité[46]. En effet, le manque de corpus oraux et le problème de leur disponibilité sous forme électronique est une critique souvent formulée à l'égard des corpus de langue générale. C'est d'autant plus vrai pour les corpus spécialisés, qui sont constitués quasi exclusivement de textes écrits (Huot 1996). D'ailleurs, un des critères fondamentaux pour la constitution de corpus spécialisés est leur publication ou leur diffusion, donc leur caractère public, qui est inévitablement lié à la forme écrite. La contrainte d'être publiés, même auprès d'un public spécialisé restreint, garantit la qualité rédactionnelle et la crédibilité des textes. En plus, le fait de n'inclure que des textes entiers augmente la fiabilité du corpus comme source de définitions (Pearson 1998). Cela dépend évidemment des objectifs de recherche. Il est à noter que les définitions et les contextes définitoires, par exemple dans les normes, sont particulièrement intéressants pour une étude sémantique s'appuyant sur l'analyse de cooccurrences, en raison des informations sémantiques précieuses qu'ils véhiculent. Les documents de notre corpus technique ont été publiés sous forme de livres (manuels et guides) et de documents électroniques (normes ISO) ou ils ont été rendus publics sur des sites Internet professionnels (revues, fiches, directives). Ils sont inclus dans leur totalité et comme ils relèvent d'un domaine technique spécialisé, les textes techniques authentiques sont factuels.

3.1.1.3 Les auteurs, le public-cible et le niveau technique

Il est préférable d'inclure un nombre important de textes différents, rédigés par plusieurs auteurs différents, tant des auteurs individuels (p.ex. revues) que des institutions ou organisations professionnelles reconnues (p.ex. normes ISO). Les auteurs doivent être reconnus comme des experts du domaine par leurs pairs et les textes inclus dans le corpus doivent être rédigés à l'origine dans la langue étudiée,

[46] Il s'agit d'articles et de documents entiers, et non pas d'extraits.

donc par des locuteurs natifs. L'original est à préférer à la traduction, car les textes traduits sont susceptibles de contenir des « expressions non-idiomatiques » (Bowker & Pearson 2002 : 52). Dans notre corpus technique, tous les textes et documents sont rédigés en français, par des francophones, à part le manuel *Fraisage à grande vitesse*, qui est une traduction. Néanmoins, nous pensons que l'inclusion de cet ouvrage, comptant à peu près 95.000 occurrences, ne pose pas de problèmes, étant donné que le traducteur, S. Torbaty, est un expert du domaine[47]. Il est l'un des deux auteurs d'un autre manuel inclus dans notre corpus technique et publié dans la même collection « Technologies d'aujourd'hui », à savoir *Rationalisation de l'Usinage très Grande Vitesse* (Cf. tableau 3.2 ci-dessous). Nous pensons dès lors que les définitions, les collocations et les expressions idiomatiques de cette œuvre traduite sont fiables et qu'elles se prêtent sans aucun problème aux analyses quantitatives de spécificités et de cooccurrences.

En ce qui concerne le public-cible et le niveau technique, un corpus spécialisé se compose généralement de textes techniques, s'adressant à un public d'experts, et de textes semi-techniques, destinés à un public de lecteurs ayant un niveau d'expertise légèrement inférieur, par exemple des initiés ou des étudiants. Les textes de vulgarisation destinés au grand public ou à un public de non-initiés, par exemple des rubriques dans des journaux, ne font pas partie d'un corpus spécialisé (Pearson 1998). Notre corpus technique se constitue de 4 sous-corpus, qui se situent à différents niveaux de vulgarisation (Cf. 3.1.1.1 la représentativité) et qui s'adressent à des publics-cibles ayant différents niveaux d'expertise technique.

3.1.1.4 Le sujet, le type de texte et le contexte communicatif

Etant donné que la recherche est limitée au domaine des machines-outils pour l'usinage des métaux, les documents s'identifient par le sujet, restreint au domaine spécialisé. La constitution détaillée de notre corpus (Cf. tableau 3.2 ci-dessous) mettra en évidence les mêmes sujets dans les différentes sources et dans les différents sous-corpus.

Différents types de textes spécialisés sont disponibles de nos jours (normes, manuels, revues, etc.). Les textes d'un corpus spécialisé se caractérisent généralement par leur caractère informatif, didactique ou normatif (Pearson 1998 ; Bowker & Pearson 2002). Les quatre sous-corpus font preuve de cette diversité de

---

[47] « Les textes traduits doivent normalement être écartés d'office (…). » « Seul un critère relatif à la qualité du traducteur (expert du domaine, membre du service de traduction d'un organisme de référence pour le domaine) peut justifier de rares exceptions » (Van Campenhoudt 2002b : 6).

types de textes, assurant la couverture conceptuelle. Le sous-corpus des normes et directives est normatif, les manuels sont didactiques et les revues et les fiches sont informatives, présentant les nouvelles technologies et découvertes (Cf. *Trametal* détectés : tableau 3.2).

Le type de texte s'inscrit aussi dans le contexte communicatif. Pearson (1998) fait la distinction entre trois contextes communicatifs : (1) entre experts, (2) entre experts et initiés, (3) entre professeurs et étudiants (Cf. visée didactique).

Revues électroniques	
Trametal www.trametal.com	**Revue technique mensuelle du travail des métaux** Archives : septembre 2000 (n°48) – mai 2002 (n°64) → Outils-coupants / machines-outils / mesure – contrôle / XAO-CNC / électroérosion / formage / équipement
Metalind www.metal-industries.com	**Mensuel de référence du travail des métaux** Articles : 1998-2001 → Technologies (formage / hydroformage / laser / jet d'eau / mesure – contrôle / profilage…)
Machpro http://www.machpro.fr/magazine/default.htm	**Revue spécialisée dans l'usinage des métaux** Articles sur les machines d'usinage (disponibles en 2002) → Fraisage / centre d'usinage / tournage / perçage, alésage / rectification / électroérosion / sciage / découpe laser, jet d'eau / machines spéciales
Fiches techniques	
Tradetec www.trametal.com	**Fiches techniques : actualités et nouveautés** Disponibles sur le site de la revue Trametal
Machout www.machine-outil.info www.machine-outil.com	**Répertoire de la machine-outil** 2000 machines avec un descriptif technique Actualités concernant la machine-outil, articles par secteur (thématique) Fiches techniques, classées par secteur (thématique) → Affûteuses / assemblage / bureaux d'études / centres d'usinage / électroérosion / fil et feuillard / filtration, aspiration, broyage / fonderie / forge / fraiseuses / jet d'eau / laser / logiciels (CAO, FAO, …) / lubrifiant / machines d'usinage spécial / maintenance / manutention / marquage / mesure, contrôle, commandes numériques / outillage, bridage / oxycoupage / perceuses / pliage / poinçonneuse, encocheuse / presse / profilés / rectifieuses / robotique / salons / sciage / sécurité / soudage / tours / traitement de surface / travail du tube

Normes et directives	
Normes http://ibn.be	**Normes catégorie E : mécanique / machine-outil** (en vente : version électronique et version papier) → Généralités : E 60 : - EN 12417 (centres d'usinage) - EN ISO 15641 (fraises pour usinage à grande vitesse) → Machines travaillant par enlèvement de métal : E 62 : - EN 12717 (perceuses) - EN 12957 (machines d'électroérosion) - EN 13128 (fraises) - EN 13218 (machines à meuler fixes)
Directiv http://normach.wtcm.be/french/directives.html	**Directives européennes (en PDF) concernant les machines** : régulation technique, champs d'application, exigences de sécurité, risques - Directive 98/37/CE du Parlement européen et du Conseil du 22 juin 1998 (45p) - La réglementation communautaire pour les machines : commentaires sur la directive 98/37/CE (1999) (255p) - Proposition de directive : COM(2000)899 (110p) - Législation belge : AR du 05-05-1995 (59p)
Guides et manuels	
Techusin	**Techniques modernes d'usinage : guide pratique** Sandvik Coromant 1997. *Techniques modernes d'usinage. Guide pratique.* Sandviken (Suède) : AB Sandvik Coromant. (868p) → Techniques d'usinage des métaux
RatUGV	**Rationalisation de l'usinage très grande vitesse** Kaufeld, M. & S. Torbaty 1999. *Rationalisation de l'Usinage très Grande Vitesse.* Boulogne : Société Française d'Editions Techniques SOFETEC. (284p) → Usinage très grande vitesse : processus d'enlèvement de matière, fraisage, tournage, broches, outils, etc.
FGV	**Fraisage à grande vitesse** Schulz, H. 1997. (traduit par S. Torbaty) *Fraisage à Grande Vitesse.* Boulogne : Société Française d'Editions Techniques SOFETEC. (343p) → Fraisage des matériaux métalliques et non-métalliques
Rectification	**Rectification des pièces de révolution** Beauchet, J. 1996. *La rectification des pièces de révolution.* Cluses : C.T.DEC. (106p) → Techniques de rectification

Tableau 3.2 Constitution détaillée du corpus technique

3.1.1.5 Critères internes et externes

Pour classer les textes, on a fait appel autant à des critères internes (linguistiques et textuels) qu'à des critères externes (extralinguistiques ou socioculturels) (Pearson 1998), ce qui correspond à la typologie interne et externe de Habert et al. (1997). Les critères externes comprennent le genre (type de texte), le mode (oral ou écrit), l'origine (l'auteur) et les objectifs (normatifs, didactiques ou informatifs). Les critères internes linguistiques portent essentiellement sur le sujet et le style.

Il est clair que les critères internes et externes reprennent les critères mentionnés précédemment. Toutefois, la distinction entre les critères internes et externes nous paraît intéressante et opérationnelle pour caractériser les particularités des corpus spécialisés.

Il est généralement admis que les critères les plus importants pour la constitution (la sélection et la compilation) d'un corpus spécialisé sont la taille et la représentativité, auxquelles Biber et al. (1998) ajoutent la diversité. Un bon corpus est représentatif par rapport aux objectifs de recherche visés, mais il est également représentatif du domaine spécialisé, de la diversité des publications et de la diversité lexicale, compte tenu du domaine. Plus un corpus est vaste, plus il a de chances d'être fiable et représentatif. Toutefois, ici aussi la prudence s'impose, car la taille ne garantit pas toujours la représentativité, notamment s'il y a un manque de diversité[48] (par exemple trop peu d'auteurs différents, trop peu de types de textes différents). La représentativité est donc plus importante que la taille.

Un corpus spécialisé se caractérise tant par son homogénéité linguistique (critère interne) que par son hétérogénéité extralinguistique (critère externe). D'une part, l'homogénéité linguistique du corpus spécialisé s'explique par le fait que les textes doivent impérativement relever du même domaine technique spécialisé (restreint) et donc porter sur les mêmes sujets techniques. D'autre part, le corpus spécialisé doit son hétérogénéité externe à la diversité des types de textes, des auteurs, des objectifs, des contextes communicatifs et éventuellement aussi des niveaux techniques. Cette diversité extralinguistique assure la bonne couverture du domaine spécialisé. Toutefois, la contrainte linguistique interne concernant le sujet reste primordiale si on veut bien délimiter le domaine et garantir la représentativité, comme le décrit Jacques (2003) ci-dessous.

[48] « Size cannot make up for a lack of diversity » (Biber et al. 1998 : 249).

> « Il serait donc préférable, pour construire les corpus en prenant l'exacte mesure de leur représentativité, d'évaluer l'homogénéité linguistique des textes, non plus seulement sur des critères externes a priori, mais aussi sur des critères internes a posteriori ». (Jacques 2003 : 66)

3.1.2 Constitution du corpus de référence

Comme nous l'avons évoqué ci-dessus, pour pouvoir déterminer les spécificités d'un corpus de langue spécialisée, il faut le comparer à un corpus de langue générale, c'est-à-dire à un corpus de référence. Un corpus de référence est conçu « pour fournir une information en profondeur sur une langue. Il vise à être suffisamment étendu pour représenter toutes les variétés pertinentes du langage et son vocabulaire caractéristique » (Habert et al. 1997 : 144). La taille du corpus de référence de langue générale sera donc plus importante que celle du corpus technique spécialisé. Généralement on adopte un rapport de 1 à 10 (Lafon 1984), pour le calcul des spécificités d'une fraction par rapport à la totalité du corpus (Cf. chapitre 4 : subdivision partie-tout).

Notre corpus de référence de langue générale est constitué d'articles journalistiques électroniques du journal *Le Monde* (de janvier à septembre 1998), disponibles sur CD-ROM. Il comprend 15.253.102 occurrences. Ainsi, le corpus technique (1,75 million d'occurrences) et le corpus de référence (15,25 millions d'occurrences) respectent le rapport de 1 à 10, c'est-à-dire que le corpus technique spécialisé représente un dixième du total de 17 millions d'occurrences.

Les critères principaux de représentativité, de diversité et de taille (Biber et al. 1998) sont également respectés dans notre corpus de référence. Il est sans aucun doute représentatif de la langue générale, puisqu'il est constitué de textes journalistiques électroniques du journal *Le Monde*. De nombreuses études de corpus ont recours au corpus du journal *Le Monde*, tant pour des expérimentations sur le français général (Guillaume & Venant 2005 ; Bourigault & Frérot 2005 ; Lamiroy & Charolles 2005 ; Habert et al. 2005) que pour confronter le corpus de langue générale à un corpus de langue spécialisée (Lemay, L'Homme & Drouin 2005). Notre corpus de référence est diversifié par la diversification thématique propre à un corpus journalistique. Il est également suffisamment étendu. D'ailleurs, il est à noter que le corpus de langue générale sert de corpus de référence pour l'analyse des spécificités et pour l'analyse des cooccurrences qui s'appuie sur la mesure de monosémie technique.

3.2 EXPLOITATION

En ce qui concerne l'exploitation et l'étude de corpus, on distingue généralement trois approches, c'est-à-dire l'approche *corpus-based*, l'approche *data-driven* et l'approche *corpus-driven* ((Tognelli-Bonelli 1994) cité par Pearson (1998)). L'approche *corpus-based* sert surtout à confirmer l'intuition du linguiste ou les théories existantes ; elle vise donc à fournir des exemples authentiques, au lieu de les construire. Puis, l'approche *data-driven* s'inscrit dans le cadre de l'apprentissage et de l'enseignement des langues. Les concordanciers et les autres outils aident les étudiants à découvrir des régularités et des règles permettant de déduire des hypothèses. Enfin, l'approche *corpus-driven* va au-delà de la sélection d'exemples confirmant une hypothèse ou théorie, car elle permet au linguiste de découvrir des phénomènes linguistiques non seulement pour valider une hypothèse, mais également et surtout pour la préciser (Pearson 1998). L'approche *corpus-driven* est adoptée principalement par la lexicographie, la terminographie et la linguistique computationnelle. Il est à noter que l'objet d'étude de la linguistique de corpus est la langue, appréhendée à travers le discours, donc à travers les réalisations effectives dans les textes du corpus (Bowker & Pearson 2002 ; Jacques 2003).

Il est clair que notre étude s'inscrit dans la perspective *corpus-driven*. D'abord, le corpus technique et le corpus de référence font l'objet d'une analyse de spécificités permettant de dresser une liste d'unités lexicales spécifiques et représentatives du corpus technique. Ensuite, le corpus technique est soumis à des analyses de cooccurrences dans le but d'étudier la sémantique de ses spécificités. Les analyses auxquelles nous procéderons vont au-delà de la sélection d'exemples et de l'analyse de listes de concordances, puisque nous visons à confirmer notre hypothèse de base, à savoir que les mots les plus spécifiques ne sont pas nécessairement les plus monosémiques. Nous allons préciser et affiner cette hypothèse par des études de corpus approfondies et détaillées, notamment par classe lexicale et par sous-corpus.

Avant d'effectuer les analyses de spécificités et de cooccurrences, les textes du corpus sont soumis à des opérations préalables de nettoyage et de catégorisation. Dans les sections suivantes, nous présenterons les différentes étapes, à savoir la préparation du corpus brut (3.2.1) ainsi que la lemmatisation et l'étiquetage des fichiers texte (3.2.2).

3.2.1 Travail de préparation du corpus brut

Pendant la compilation et la sélection des textes du corpus technique, l'acquisition et la préparation du matériel dépendent de l'accessibilité et de la disponibilité des sources. Dans notre corpus technique, il s'agit de documents *.pdf téléchargés sur Internet, de documents Internet copiés (HTML ou non) et de documents numérisés.

Pour le corpus de référence, nous avons utilisé le CD-ROM du journal *Le Monde*. Les textes sont sauvegardés sous le format de traitement de texte *.txt (3.2.1.1) et font l'objet d'opérations de nettoyage et de correction (3.2.1.2).

3.2.1.1 Fichiers texte

Comme le format *.txt ne prévoit pas de mise en page (gras, italique, etc.), c'est le meilleur format pour sauvegarder des textes téléchargés, copiés ou numérisés, parce que les fichiers *.txt sont peu volumineux et dès lors faciles à stocker et à manipuler, étant donné qu'un corpus comprend généralement plusieurs millions de mots.

Les directives, les normes et les numéros de la revue *Trametal* sont disponibles et téléchargeables au format *.pdf, ce qui permet de sauvegarder les textes sous le format de traitement de texte *.rtf ou *.txt. D'abord, les documents ont été sauvegardés sous *.rtf (Word), avec maintien du formatage, mais sans graphismes, afin de contrôler et de corriger la conversion des caractères. Ensuite, les documents *.rtf ont été réunis dans un fichier *.txt par source et pour la revue Trametal, dans deux fichiers *.txt, en raison de la taille trop importante pour la lemmatisation ultérieure (Cf. 3.2.2.1). Les documents des autres revues électroniques et des fiches techniques ont été téléchargés et copiés dans un fichier *.txt par source. Troisièmement, les guides et manuels ont été numérisés à l'aide du logiciel d'OCR (*Optical Character Recognition*) OmniPage Pro 11. Lors de la numérisation, l'application OCR de reconnaissance optique de caractères permet de reconnaître plusieurs colonnes, d'ignorer les graphismes et de sauvegarder le texte sous le format *.txt. Dans les textes numérisés, quelques petites fautes de reconnaissance ont été corrigées. Il est à noter que la numérisation (manuelle) de quatre livres représente un travail fastidieux et de longue haleine, mais efficace en raison de la bonne qualité du résultat. Au total, les douze fichiers *.txt du corpus technique représentent environ 10 Mo de texte plein.

Le corpus de référence a été téléchargé du CD-ROM du journal *Le Monde* et sauvegardé en plusieurs fichiers *.txt. Tous les fichiers texte du corpus de référence constituent un corpus sous le format de traitement de texte *.txt d'environ 89 Mo.

Signalons tout de même que les graphismes et les photos ont été enlevés, étant donné les objectifs essentiellement linguistiques de notre étude. Si besoin en est, on pourra toujours recourir aux documents d'origine et consulter les images et données visuelles. En plus, toutes les informations originales permettant d'identifier les textes ont été maintenues (numérotation des pages, indication de la source, etc.). Cependant, le corpus technique est très hétérogène en ce qui concerne la mise en page, étant donné qu'il consiste en onze sources différentes et autant d'indications de mise en page différentes.

Il est à noter également que généralement, les documents électroniques des corpus sont balisés, ce qui facilite l'exploitation, le dépouillement, le partage et l'échange. Les balises de la recommandation T.E.I. (*Text Encoding Initiative*) fournissent des informations sur les documents, telles que le titre et les paragraphes (Van Campenhoudt 2002b ; Bowker & Pearson 2002), et permettent notamment l'alignement par utilisation d'identifiants parallèles. Le langage de balisage le plus courant est le codage XML (*eXtensible Markup Language*), utilisé pour structurer et échanger des ressources textuelles et pour permettre une séparation fond/forme. La syntaxe XML consiste en une chaîne de caractères, encadrée par des chevrons ouvrant et fermant, par exemple <header>. A chaque balise ouvrante, telle que <text> correspond une balise fermante, par exemple </text>.

En raison du caractère hétérogène de la mise en page des documents de notre corpus technique, nous n'avons pas eu recours au balisage XML de la TEI. La source du document est l'indication la plus importante pour nos objectifs de recherche actuels. Un balisage complet du corpus constituerait une recherche en soi et dépasserait les limites de cette étude, qui se veut avant tout une étude sémantique quantitative. Cela dit, nous n'excluons pas un balisage ultérieur plus fin et plus détaillé, en vue d'une homogénéisation plus poussée des documents du corpus et devant permettre de réaliser d'autres objectifs de recherche.

3.2.1.2 Opérations de nettoyage avant lemmatisation

Les douze fichiers *.txt des différentes sources du corpus technique ont fait l'objet de plusieurs opérations de nettoyage et de correction[49], consistant notamment à corriger les fautes de frappe et les fautes d'orthographe. Nous avons également procédé à d'autres opérations de nettoyage, au niveau *.txt, c'est-à-dire avant la lemmatisation[50] des textes, afin d'éviter que les erreurs des formes graphiques ne se (re)produisent dans les lemmes. Toutes les opérations de nettoyage et de correction mentionnées ci-dessous sont décrites en détail dans un document en annexe (Cf. annexe 1).

[49] « La phase initiale de nettoyage et d'homogénéisation des textes collectés sous forme électronique est une étape souvent sous-estimée, alors qu'elle est cruciale » (Habert et al. 1997 : 141).

[50] Notons que les fautes à corriger ont été découvertes après une première lemmatisation. Au moment d'importer les lemmes et les formes graphiques dans Access, les opérations de tri et de recherche automatique ont permis de relever ces fautes de frappe. Après correction, nous avons procédé à une deuxième lemmatisation, définitive.

Notons que l'éditeur de texte Textpad permet l'utilisation d'expressions régulières, ce qui facilite considérablement la correction. Etant donné que ces fautes sont essentiellement dues à la numérisation et à la conversion des documents *.pdf, elles ont été corrigées dans le corpus technique, mais pas dans le corpus de référence.

1) Correction de mots composés coupés en fin de ligne (avec saut de ligne (\n) et où le deuxième élément se trouve à la ligne suivante), ce qui donne lieu à une lemmatisation fautive, par exemple : *porte-***(\n)***outil*

2) Correction de mots avec trait d'union pour des raisons typographiques (division intentionnelle), par exemple *automatique-ment*

3) Correction de fautes de frappe (mots avec et sans trait d'union), par exemple *celuici, semif-inition*

4) Correction pour *kN* et *fZ* (lemmatisation : *kn-et, kn-avec, fz-),* …), où l'ajout d'un saut de ligne (\n) permet d'éviter une erreur de lemmatisation

3.2.2 Lemmatisation et étiquetage du corpus

Les étapes suivantes consistent à lemmatiser et à étiqueter les fichiers texte (3.2.2.1) et à nettoyer les fichiers lemmatisés (3.2.2.2). La lemmatisation permet de rattacher un mot à sa forme canonique : les adjectifs sont ainsi ramenés à la forme du masculin singulier, les substantifs sont ramenés à la forme du singulier, les verbes sont ramenés à l'infinitif (*dangereuses* → *dangereux, machines* → *machine, permettent* → *permettre*). L'étiquetage morphosyntaxique (ou la catégorisation) revient à identifier la catégorie morphosyntaxique d'une forme graphique, en contexte. Pour la détermination des spécificités ou mots-clés (Cf. chapitre 4), nous avons besoin des lemmes de toutes les formes graphiques du corpus technique et du corpus de référence. En effet, les spécificités sont déterminées au niveau des lemmes (p.ex. *machine*) et non pas au niveau des formes graphiques, où l'on aurait pour les substantifs par exemple *machine* et *machines* et pour les verbes toutes les formes conjuguées (p.ex. *permet, permettent, permettra, permettant* du lemme *permettre*). L'étiquetage morphosyntaxique permet aussi d'identifier la classe lexicale, indispensable notamment à la subdivision des spécificités par classe lexicale.

3.2.2.1 Fichiers lemmatisés

Les fichiers texte du corpus technique et du corpus de référence ont été lemmatisés avec le logiciel Cordial 7 Analyseur[51], qui « offre une lemmatisation et un étiquetage

[51] Synapse Développement Editeur de logiciels : http://www.synapse-fr.com/.

morphosyntaxique d'une exactitude satisfaisante » (Audibert 2003 : 36). Le logiciel Cordial accorde des codes pour marquer la catégorie morpho-syntaxique, par exemple 0 pour l'adjectif masculin singulier, 24 pour le substantif masculin singulier et 25 pour le substantif masculin pluriel. Sous forme de code, le logiciel ajoute donc des informations morphologiques supplémentaires, telles que la distinction entre singulier / pluriel pour les adjectifs et les noms, masculin / féminin pour les adjectifs, temps / mode / personne pour les verbes.

Les fichiers générés par le logiciel Cordial (avec extension *.cnr) se composent de trois colonnes, séparées par des tabulations et avec un mot par ligne (Cf. tableau 3.3 ci-dessous) : (1) la forme fléchie ou forme graphique, (2) le lemme ou forme canonique et (3) le code Cordial, comparable à un POS-tag (*Part-Of-Speech*) indiquant la classe lexicale. Ainsi, douze fichiers *.cnr ont été générés pour les douze fichiers *.txt correspondants, équivalents à 28,8 Mo. Pour le corpus de référence, les fichiers lemmatisés représentent environ 249 Mo. Les fichiers *.cnr, ou fichiers lemmatisés, sont parfaitement lisibles par un éditeur de texte, tel que Textpad.

Pour ne pas surcharger inutilement les documents étiquetés, nous avons uniquement procédé à la lemmatisation et à l'étiquetage morphosyntaxique (*tagging*), qui étaient indispensables aux analyses envisagées. Nous n'avons pas procédé à une analyse syntaxique complète (*parsing*).

```
Ce      ce      09
manuel  manuel  24
passe   passer  103
en      en      23
revue   revue   26
les     le      16
techniques      technique       27
modernes        moderne         07
d'      de      23
usinage         usinage         24
des     de      16
métaux  métal   25
```

Tableau 3.3 Exemple de texte étiqueté par Cordial

3.2.2.2 Opérations de nettoyage après lemmatisation

Les fichiers lemmatisés ont fait l'objet d'un nettoyage, qui a consisté à vérifier et à corriger des erreurs de lemmatisation et à faire quelques regroupements. Cette opération a été effectuée pour la version lemmatisée tant du corpus technique que du corpus de référence. Elle est expliquée en détail dans les documents en annexe, pour le corpus technique (Cf. annexe 2) et pour le corpus de référence (Cf. annexe 3).

Voici les opérations de nettoyage et de correction après lemmatisation :

1) Vérification et correction des erreurs de lemmatisation :

   a) erreurs de lemmatisation, par exemple *machines-outil* → *machine-outil*

   b) erreurs subsistantes : *kn-* → *kn* et *fz-* → *fz* (pendant la lemmatisation : ajout d'un tiret au lemme)

2) Regroupements : les lemmes à double graphie (par exemple avec et sans majuscule) ont été regroupés sous le lemme le plus fréquent

   a) lemmes avec majuscule et avec minuscule, par exemple *Fig.* et *fig.*

   b) lemmes avec point et sans point, par exemple *etc.* et *etc* ou *Fig.* et *Fig*

   c) lemmes avec trait d'union et sans trait d'union (les deux variantes sont attestées et possibles) (Cf. annexe 1), par exemple *ultra-fin* et *ultrafin*

3) Opérations de nettoyage et de correction supplémentaires

Les opérations de regroupement des lemmes à double graphie permettent de regrouper sous un seul lemme des lemmes qui s'écrivent différemment, mais qui devront être considérés comme un seul lemme au moment de générer la liste des spécificités. Ainsi, dans les fichiers lemmatisés (*.cnr), à côté de la première colonne des formes graphiques *Fig.* et *Fig* (avec et sans point), on indiquera, dans la deuxième colonne des lemmes, le lemme le plus fréquent, à savoir *Fig* (sans point). Par conséquent, au moment de dresser la liste de fréquence des lemmes (Cf. 3.3) qui sert de fichier d'entrée pour la liste des spécificités, on aura uniquement le lemme *Fig* (sans point) et on évitera de retrouver dans la liste des spécificités les deux graphies comme deux spécificités différentes. Il en va de même pour les lemmes avec et sans majuscule et pour les lemmes avec et sans trait d'union.

3.3 PRÉPARATION AUX ANALYSES

Avant de passer aux chapitres méthodologiques proprement dits, consacrés aux spécificités (chapitre 4) et aux cooccurrences (chapitre 5), nous nous proposons de décrire la préparation des documents et des listes indispensables aux analyses des spécificités et des cooccurrences.

Le dépouillement et les analyses de corpus se font généralement à l'aide d'outils et de logiciels de dépouillement, notamment pour la génération de concordances et de

listes de fréquence. Mentionnons par exemple les logiciels WordCruncher[52], WordSmith[53], Lexico3[54] et Abundantia Verborum[55]. Il est clair que les outils utilisés dépendent des objectifs de recherche ainsi que des exigences spécifiques de l'analyse. Nous avons remédié aux lacunes des logiciels mentionnés en élaborant des scripts en Python[56] pour réaliser certaines analyses.

Pour les analyses des spécificités, nous aurons besoin d'une liste de fréquence des lemmes du corpus technique ainsi que du corpus de référence. La première section (3.3.1) sera consacrée aux listes de fréquence, générées à l'aide d'un script en Python. Etant donné que les mots grammaticaux et les noms propres seront enlevés de la liste des spécificités, nous en dresserons également la liste (3.3.2). Il est évident que les codes Cordial des fichiers lemmatisés, qui indiquent la classe lexicale, seront très utiles à cet effet. Dans la dernière section, nous procéderons finalement à une comparaison quantitative entre le corpus technique et le corpus de référence (3.3.3), aussi bien en termes de lemmes que de formes graphiques.

3.3.1 Listes de fréquence du corpus technique et du corpus de référence

L'opération de base de la linguistique de corpus, avant tout dépouillement du corpus, consiste à dresser des listes de fréquence. Une liste de fréquence donne tous les mots d'un corpus avec leur fréquence d'occurrence. Nous dressons non seulement la liste de fréquence des formes graphiques, mais également la liste de fréquence des lemmes, tant pour le corpus technique que pour le corpus de référence. En effet, les deux listes de fréquence des lemmes sont requises pour l'analyse des spécificités (Cf. chapitre 4). Les deux listes de fréquence des formes graphiques (formes fléchies) sont indispensables en outre pour l'analyse des cooccurrences, plus particulièrement pour le calcul de la mesure de recoupement technique (Cf. chapitre 6). En plus, toutes ces listes de fréquence permettent aussi

---

[52] WordCruncher : http:/www.wordcruncher.com.

[53] WordSmith Tools version 3 : http://www.lexically.net/wordsmith/ et http://www.oup.com.

[54] Lexico3 : SYLED – CLA2T, Paris3 :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/.

[55] Abundantia Verborum : http://wwwling.arts.kuleuven.be/genling/abundant/obtain/.

[56] Nous nous sommes basée sur une bibliothèque de scripts existants, que nous avons adaptés et enrichis.

une comparaison quantitative entre le corpus technique et le corpus de référence (Cf. 3.3.3).

Par conséquent, quatre listes de fréquence ont été générées, à savoir :

1) une liste de fréquence des lemmes du corpus technique

2) une liste de fréquence des lemmes du corpus de référence

3) une liste de fréquence des formes graphiques du corpus technique

4) une liste de fréquence des formes graphiques du corpus de référence

Un script en Python parcourt les différents fichiers lemmatisés par corpus (soit la colonne des lemmes, soit la colonne des formes graphiques) et produit en sortie un fichier texte avec les mots (soit lemmes, soit formes graphiques) et leur fréquence, indiquée devant le mot (fréquence TAB mot) (Cf. tableau 3.4). Les détails de la génération des listes de fréquence sont expliqués dans le document en annexe (Cf. annexe 4). Notons que les mots composés avec trait d'union (-) et avec barre oblique (/) sont reconnus et lemmatisés en tant que tels par Cordial, comme le montre le tableau 3.4 ci-dessous, pour la liste de fréquence des lemmes du corpus technique.

```
5       spacieux
1       inspecteur
237     affaire
13      coulisser
1       cristallisation
19      intrinsèque
323     porte-outil
1       ensachage
1       encliquetage
2       m/min
2       4.1.3
127     accroissement
76      machiner
101     filtration
2       unanimité
7       anglo-saxon
803     protection
```

Tableau 3.4 Extrait de la liste de fréquence des lemmes du corpus technique

3.3.2 Listes de mots grammaticaux et de noms propres

L'analyse des spécificités s'appuie donc sur les deux listes de fréquence des lemmes du corpus technique et du corpus de référence, dans le but de dresser la liste des spécificités ou mots-clés (Cf. chapitre 4). Toutefois, cette liste de spécificités devra

encore être filtrée parce que seuls les mots lexicaux ou les mots « pleins », à savoir les substantifs, adjectifs, verbes et adverbes, seront analysés, contrairement aux mots grammaticaux ou mots fonctionnels (*function words* ou *stopwords*), qui sont sémantiquement « vides ». Ces derniers devront être supprimés de la liste des spécificités, tout comme les noms propres d'ailleurs. En effet, ils ne seront pas intégrés dans la liste des spécificités, car l'analyse portera uniquement sur les spécificités lexicales du corpus technique.

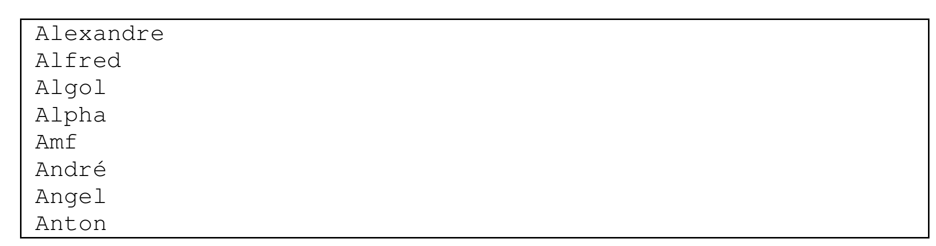
En guise de préparation aux analyses, nous procédons dès lors à l'établissement des listes de mots grammaticaux et de noms propres. A cet effet, les codes Cordial des fichiers lemmatisés s'avèrent particulièrement utiles, car ils indiquent non seulement la catégorie grammaticale des lemmes, mais ils font également la distinction entre les noms communs et les noms propres. Pour le français, il existe quelques listes de mots grammaticaux[57] (*stopword lists*). Pour l'anglais par contre, ces listes sont plus nombreuses. Les listes de mots grammaticaux pour le français dépendent en partie du corpus dont ils ont été extraits. Par conséquent, nous avons décidé de ne pas utiliser ces listes, étant donné que nous disposons de la catégorisation par Cordial. Les codes Cordial s'y prêtent bien, à condition de prévoir quelques vérifications supplémentaires. Les opérations de génération des listes de mots grammaticaux et de noms propres sont décrites en détail dans le document en annexe (Cf. annexe 5).

Les mots grammaticaux et les noms propres sont relevés au niveau des lemmes et non pas au niveau des formes graphiques (ou formes fléchies), parce que le but est de filtrer la liste des spécificités, qui se présentent sous forme de lemmes. Les éléments à retrouver et à supprimer doivent être identiques formellement pour les filtrer de façon automatisée à l'aide d'un script. La liste des mots grammaticaux comprend 448 mots grammaticaux, catégorisés ainsi par Cordial. La liste des noms propres a recensé 7200 noms propres. Ces deux listes, qui contiennent un mot par ligne, feront l'objet d'opérations ultérieures de filtrage (Cf. tableaux 3.5 et 3.6).

```
au-dessus
auparavant
auprès
auquel
aussi
aussitôt
autant
```

Tableau 3.5 Extrait de la liste des mots grammaticaux du corpus technique

[57] Signalons à ce propos la liste de Véronis : antidictionnaire (*stoplist*), disponible sur : http://www.up.univ-mrs.fr/~veronis/donnees/index.html.



```
Alexandre
Alfred
Algol
Alpha
Amf
André
Angel
Anton
```

Tableau 3.6 Extrait de la liste des noms propres du corpus technique

Même si l'annexe 5 détaille toutes les opérations effectuées pour dresser les listes de mots grammaticaux (448) et de noms propres (7200), nous tenons à commenter et à justifier certaines décisions concernant les lemmes à inclure ou à exclure, tant (1) pour les mots grammaticaux que (2) pour les noms propres.

1) Mots grammaticaux

La liste des mots grammaticaux contient non seulement des mots purement grammaticaux, appartenant aux classes fermées des mots fonctionnels ou des mots sémantiquement « vides », tels que des pronoms et des conjonctions. Elle contient également des adverbes grammaticaux et des auxiliaires, dans la mesure où ils sont grammaticalisés (Lamiroy 1998 ; Lamiroy & Charolles 2004 et 2005). Afin de relever les adverbes grammaticaux, nous avons adopté un critère formel, à savoir la terminaison, puisque les adverbes lexicaux se terminent majoritairement par *–ment*, par exemple *hydrauliquement*. Les adverbes grammaticaux, par contre, ne se terminent pas nécessairement par *–ment*, par exemple *surtout, ensuite, désormais*.

Cependant, parmi les adverbes en *–ment*, il y a encore des adverbes conjonctifs (Piot 1996 ; Lamiroy & Charolles 2004), à savoir *également, exclusivement, notamment, particulièrement, seulement, simplement, singulièrement, spécialement, uniquement*. Piot considère les adverbes *seulement, simplement, exclusivement et uniquement* comme des items restrictifs, permettant « la conjonction de deux constituants » (Piot 1996 : 343). *Seulement* et *simplement* peuvent aussi « intervenir comme joncteurs entre 2 phrases prises dans leur ensemble » (Piot 1996 : 343). C'est le cas lorsqu'ils se trouvent en tête de phrase, position dans laquelle ils adoptent « des propriétés conjonctives très proches de celles de *mais* » (Lamiroy & Charolles 2005 : 117). Ils passent donc de l'emploi intraprédicatif et restrictif à un emploi conjonctif et oppositif. *Egalement* signifie une addition et permet la « conjonction de deux phrases ou de deux constituants entre deux phrases parallèles » (Piot 1966 : 344). Finalement, *notamment, particulièrement, singulièrement* et *spécialement* « indiquent sémantiquement une emphase ou focalisation », ce qui justifie leur emploi « comme ajout nominal ou verbal ou comme joncteur entre deux phrases entières (de contenu non parallèle) » (Piot 1996 : 345). Il en va de même pour *autrement*, adverbe de manière et connecteur (Lamiroy & Charolles 2005).

Bien que ces adverbes puissent servir de conjonctions, leur emploi adverbial[58] est également attesté ou même plus fréquent que leur emploi conjonctif[59] [60] (Lamiroy & Charolles 2004 et 2005). Afin de connaître l'emploi prédominant de ces adverbes dans le corpus technique, nous avons mené une expérimentation sur 100 occurrences aléatoires de ces adverbes (Cf. annexe 5 : expérimentation adverbes). Il en ressort que l'emploi adverbial (intraprédicatif) est très largement prédominant (99%) et que l'emploi conjonctif (extraprédicatif) est tout à fait marginal (1%). Dès lors, ces adverbes conjonctifs en *–ment* ne feront pas partie de la liste des mots grammaticaux. Ils seront intégrés dans la liste des mots lexicaux. Il est toutefois à noter que les mots grammaticaux, y compris les adverbes qui ont également un code Cordial en tant que mots lexicaux (adjectifs ou substantifs) sont exclus de la liste des mots grammaticaux. Ils seront donc analysés au même titre que les mots lexicaux.

2) Noms propres

Pour la liste des noms propres, nous avons également procédé à des vérifications et à des décisions d'inclusion et d'exclusion. En effet, certains lemmes ont des codes erronés et ne devraient pas faire partie de la liste de noms propres. La fréquence par code Cordial est une indication fiable de ce problème, que l'on peut détecter avec la liste des doublons avec au moins un code de nom propre (75) (Cf. tableau 3.7). Ainsi les abréviations et sigles, qui ont reçu le code des noms propres à cause de la majuscule, ont aussi été supprimés de la liste des noms propres. Nous avons décidé de les intégrer dans les analyses sémantiques détaillées, en raison de leur statut particulier dans le corpus technique, par exemple *Cfao*, qui signifie « conception et fabrication assistée par ordinateur ».

lemme	code Cordial	fréquence par code Cordial
aléser	100	51
aléser	75	1
aléseuse-fraiseuse	26	3
aléseuse-fraiseuse	27	4
aléseuse-fraiseuse	75	2

Tableau 3.7 Doublons avec au moins un code de nom propre

[58] *J'ai seulement 20 euros.*

[59] *J'ai promis d'assister à la conférence, seulement, je n'ai pas le temps.*

[60] D'après l'étude de corpus de Lamiroy & Charolles (2004 et 2005), l'emploi adverbial est largement prédominant (96%) dans un corpus journalistique.

3.3.3 Comparaison : corpus technique – corpus de référence

Nous allons procéder à une comparaison quantitative du corpus technique et du corpus de référence, en termes de formes graphiques (ou formes fléchies) et en termes de lemmes (ou formes canoniques). L'étendue d'un corpus s'évalue généralement par le nombre total d'occurrences, c'est-à-dire par les formes graphiques ou fléchies apparaissant dans le corpus. L'étendue ou la taille d'un corpus équivaut donc au nombre total de formes graphiques (*tokens*) (1) (Cf. tableau 3.8), même si celles-ci sont récurrentes. Par contre, si les répétitions ne sont pas prises en considération, on calcule le nombre de formes graphiques différentes (*types*) (2). La version lemmatisée du corpus permet de déterminer également le nombre de lemmes différents (ou le nombre de formes canoniques différentes) (4). Le nombre total de lemmes (3) est égal au nombre total de formes graphiques (1) (Cf. tableau 3.8). Les formes graphiques et lemmes indiqués dans le tableau ci-dessous ne comprennent pas de signes de ponctuation[61], ni au niveau des *tokens*, ni au niveau des *types*.

Le rapport entre le nombre de formes graphiques différentes (2) et le nombre total de formes graphiques (1), appelé le *Type-Token Ratio* ou TTR[62] (5), permet de mesurer la richesse lexicale du corpus ou la diversité de son vocabulaire (Manning & Schütze 2002). Plus le TTR est élevé, plus il y a de formes différentes dans le corpus. Il est à noter que le TTR est toujours calculé pour un corpus donné. Dès lors, le TTR est influencé par les sujets traités dans le corpus et par la longueur du corpus. Des sujets hétérogènes entraînent effectivement plus de formes différentes (et plus de lemmes différents) et donc un TTR plus élevé. En plus, dans un texte plus long, les mots ont plus de chances d'être répétés, ce qui pourrait se traduire par un TTR plus faible. Pour une comparaison valable, il faut dès lors normaliser la longueur des textes ou la taille des corpus (Manning & Schütze 2002 ; Van Campenhoudt 2002b) : soit en comparant des corpus de taille identique, soit en comparant les TTR standardisés[63], en « calculant la mesure TTR pour des fenêtres de mille mots »

---

[61] Etant donné que les signes de ponctuation sont pourvus des codes Cordial de 201 à 209, il est facile de les éliminer des opérations de décompte pour le corpus technique (180.737) et pour le corpus de référence (2.013.732). Par rapport au total des signes (mots et signes de ponctuation), les signes de ponctuation représentent 10,7% du corpus technique et 8,6% du corpus de référence, ce qui confirme la particularité de la langue spécialisée du domaine technique en ce qui concerne le suremploi de signes de ponctuation.

[62] Formule généralement adoptée (Van Campenhoudt 2002b ; WordSmith Tools WordList) : *(nombre de formes graphiques différentes *100) / nombre total de formes graphiques.*

[63] Il est à noter que les TTR standardisés ne permettent pas de résoudre le problème de la diversité des sujets traités.

(Manning & Schütze 2002 : 22). A cet effet, nous allons comparer le corpus technique à un échantillon aléatoire du corpus de référence de taille comparable (1,7 million d'occurrences) (Cf. tableau 3.9).

	corpus technique entier	corpus de référence entier
(1) Nombre total de formes graphiques (*tokens*)	1.751.800	15.253.102
(2) Nombre de formes graphiques différentes (*types*)	47.636	254.061
(3) Nombre total de lemmes (*tokens*)	1.751.800	15.253.102
(4) Nombre de lemmes différents (*types*)	29.426	152.128
(5) TTR formes graphiques	**2,71926019**	1,665634964
(6) TTR lemmes	**1,679757963**	0,997357783
(7) *Token-Type Ratio* : formes graph.	36,7747082	60,0371643
(8) *Token-Type Ratio* : lemmes	59,53238633	100,2649216
(9) *Types* formes graphiques / lemmes	1,618840481	1,670047592

Tableau 3.8 Lemmes et formes graphiques : corpus technique – corpus de référence

En général, le TTR est calculé pour les formes graphiques (fléchies), mais le calcul du TTR est également envisageable au niveau des lemmes (6), ce qui revient à diviser le nombre de lemmes différents par le nombre total de lemmes (qui est égal au nombre total de formes graphiques). Bien entendu, le rapport TTR des lemmes est plus faible que le TTR des formes graphiques, parce que chaque forme a été ramenée à sa forme canonique. Ainsi, deux formes graphiques (*types*) telles que *machine* et *machines* relèvent du même lemme *machine* et toutes les formes conjuguées (*types*) d'un verbe par exemple relèvent du même lemme, qui est l'infinitif. Pour des langues à forte flexion, comme le français[64], le *type* ne correspond pas du tout au lemme (Van Campenhoudt 2002b), ce qui se reflète d'ailleurs dans le rapport des types des formes graphiques divisés par les types des lemmes (9). Le rapport de 1,6 indique qu'il y a environ 1,6 forme graphique par lemme. Le TTR des lemmes donne des résultats plus précis et plus fiables en matière de richesse lexicale (Van Campenhoudt 2002b).

Le tableau 3.8 suggère que le corpus technique serait plus diversifié et plus riche lexicalement, pour les lemmes et pour les formes graphiques : les TTR (5) et (6) du corpus technique (2,7 et 1,6) étant supérieurs à ceux du corpus de référence (1,6 et

[64] C'est relatif. Le français est beaucoup moins flexionnel que les autres langues romanes, mais plus flexionnel que l'anglais.

0,9). Comme le rapport TTR dépend fortement de la taille des corpus considérés, nous proposons une telle comparaison des rapports de TTR pour un échantillon aléatoire du corpus de référence, de taille comparable à la taille du corpus technique (Cf. tableau 3.9). Cette comparaison montre effectivement le contraire : le TTR des formes graphiques (5) et le TTR des lemmes (6) sont plus élevés dans l'échantillon du corpus de référence (4,7 et 2,8)[65] que dans le corpus technique (2,7 et 1,6). Pour le même nombre total de formes et de lemmes, approximativement, le corpus de référence recense donc beaucoup plus de formes différentes (2) et de lemmes différents (4) que le corpus technique.

	corpus technique entier	corpus de référence échantillon
(1) Nombre total de formes graphiques (*tokens*)	1.751.800	1.747.452
(2) Nombre de formes graphiques différentes (*types*)	47.636	82.924
(3) Nombre total de lemmes (*tokens*)	1.751.800	1.747.452
(4) Nombre de lemmes différents (*types*)	29.426	49.174
(5) TTR formes graphiques	2,71926019	**4,745423623**
(6) TTR lemmes	1,679757963	**2,8140401**
(7) *Token-Type Ratio* : formes graph.	36,7747082	21,07293425
(8) *Token-Type Ratio* : lemmes	59,53238633	35,53609631
(9) *Types* formes graphiques / lemmes	1,618840481	1,686338309

Tableau 3.9 Lemmes et formes graphiques : corpus technique – échantillon du corpus de référence

La richesse lexicale ou la diversité lexicale du corpus de référence s'explique principalement par l'hétérogénéité thématique des textes journalistiques dont il est constitué. Le corpus technique, au contraire, se caractérise par l'homogénéité thématique et dès lors par la récurrence plus importante des formes, étant donné que les documents du corpus technique relèvent tous d'un domaine spécialisé, restreint par définition (Cf. 3.1.1.4). Cette répétition ou récurrence s'exprime par le rapport inverse du TTR, à savoir le *Token-Type Ratio*, qui indique la fréquence d'occurrence moyenne par forme graphique (7) ou par lemme (8).

[65] Ces chiffres se vérifient pour un autre échantillon du corpus de référence de taille comparable (4,6 et 2,8 respectivement).

Dans le corpus technique, les formes graphiques figurent 36,7 fois en moyenne[66]. Dans l'échantillon du corpus de référence de langue générale, la fréquence moyenne des formes graphiques est moins élevée : elle n'est que de 21. A titre d'information, la fréquence moyenne des formes dans le corpus de référence entier est de 60. Si le texte est plus long, les formes ont clairement plus de chances d'être répétées, étant donné le nombre limité de formes (et certainement de lemmes) dans la langue. Le *Token-Type Ratio* des lemmes (8) confirme ainsi la diversité lexicale du corpus de référence. Les lemmes sont répétés en moyenne 35,5 fois dans l'échantillon du corpus de référence de langue générale, tandis que la récurrence moyenne des lemmes dans le corpus technique est de 59,5. En dépit de cette différence en matière de diversité lexicale, les deux corpus se caractérisent par un rapport comparable (1,6) entre le nombre de formes graphiques différentes et le nombre de lemmes différents (9).

[66] Dans un corpus, très peu de mots sont très fréquents (p.ex. les articles *de* et *le* ou les substantifs *machine* et *usinage* dans notre corpus technique) et la plupart des mots sont très peu fréquents, un nombre important étant même des hapax. Si les mots d'un corpus sont classés par ordre décroissant de fréquence, la fréquence des mots est inversement proportionnelle à leur rang, selon la loi de Zipf (Manning & Schütze 2002).

Chapitre 4
Analyse des spécificités

Le quatrième chapitre vise à expliciter le premier axe méthodologique de cette étude, à savoir l'axe des spécificités. Le deuxième axe méthodologique des cooccurrences fera l'objet du cinquième chapitre. En effet, avant de procéder à l'analyse proprement dite, il faudra déterminer la sélection de mots sur lesquels l'analyse sémantique sera basée. Il faudra donc déterminer les mots les plus spécifiques du corpus technique, c'est-à-dire les mots-clés ou les spécificités[67]. A cet effet, le corpus technique de langue spécialisée sera comparé à un corpus de référence de langue générale, permettant d'identifier les unités lexicales (simples) spécifiques et représentatives du corpus technique. Dans cette étude, nous nous limitons aux unités simples, dont les lemmes dans le corpus technique seront comparés aux lemmes formellement identiques dans le corpus de référence. Rappelons que si les unités simples les plus spécifiques du corpus technique s'avèrent les plus monosémiques, la thèse monosémiste traditionnelle se verra confirmée. Dans ce chapitre, nous expliquerons la méthodologie à laquelle nous recourons pour identifier les spécificités et pour déterminer leur degré de spécificité.

Comme il existe plusieurs méthodes statistiques pour déterminer les spécificités, nous expliquerons d'abord les deux approches méthodologiques principales, à savoir le calcul des spécificités et la méthode des mots-clés (4.1). Dans la deuxième partie, nous nous proposons de comparer trois outils pour l'identification des spécificités qui sont basés sur ces deux approches (4.2). Nous terminerons par la justification de la méthodologie adoptée dans notre étude, à savoir la méthode des mots-clés (4.3).

---

[67] Nous adoptons le terme « spécificités » pour désigner les mots les plus spécifiques et caractéristiques du corpus technique, indépendamment de la méthode utilisée (calcul des spécificités vs. méthode des mots-clés).

4.1 DEUX APPROCHES MÉTHODOLOGIQUES

Les recherches en langue spécialisée prennent souvent comme point de départ l'identification des spécificités, c'est-à-dire des mots spécifiques qui caractérisent le corpus de langue spécialisée et qui le différencient d'un corpus de langue générale. Soulignons d'emblée que les spécificités ne sont pas les mots les plus fréquents[68] du corpus de langue spécialisée, mais les mots les plus caractéristiques et les plus représentatifs[69]. D'ailleurs, les mots les plus fréquents sont surtout des mots grammaticaux ou fonctionnels, tels que *le, un, à, avec*, qui sont écartés de l'analyse. D'un point de vue relatif, les spécificités apparaissent plus fréquemment dans le corpus de langue spécialisée que dans un corpus de référence de langue générale, et cela de manière significative.

Afin de déterminer les spécificités, les fréquences dans le corpus spécialisé sont comparées aux fréquences dans le corpus de référence de langue générale, compte tenu de la taille des deux corpus. Cela revient à comparer la fréquence observée dans le corpus spécialisé à la fréquence attendue dans le corpus spécialisé à partir des observations dans le corpus de référence. S'il y a une différence entre la fréquence observée et la fréquence attendue et si cette différence est statistiquement significative, elle permet d'identifier les spécificités (Bertels 2005). A cette fin, deux approches méthodologiques sont envisageables, d'une part, le calcul des spécificités (4.1.1) et, d'autre part, la méthode des mots-clés ou *Keywords Method* (4.1.2).

Les deux méthodologies aboutissent en gros à des résultats similaires, à savoir une liste de mots pourvus d'une mesure statistique indiquant leur degré de spécificité. Les différences les plus importantes résident dans la méthodologie et la statistique sous-jacentes, décrites pour les deux approches dans les deux sections méthodologiques suivantes (4.1.1 et 4.1.2).

---

[68] Il convient de noter que les mots les plus spécifiques sont également très fréquents puisque les mots peu fréquents dans le corpus technique ne seraient pas représentatifs de ce corpus et ne figureraient pas parmi les mots les plus spécifiques.

[69] A titre d'expérimentation, les 4717 mots les plus spécifiques du corpus technique ont été comparés aux 4757 mots les plus fréquents du corpus technique (ayant une fréquence absolue dans le corpus technique $\geq$ 18). Le recoupement est important : parmi les 4717 mots spécifiques, 2548 mots appartiennent à la liste des 4757 mots les plus fréquents (54%), les autres 2169 spécificités étant moins fréquentes (< 18).

4.1.1 Le calcul des spécificités

La première approche méthodologique pour l'identification des spécificités est le calcul des spécificités (Lafon 1984 ; Müller 1992[70]). Du point de vue méthodologique, le calcul des spécificités procède par comparaison partie-tout. Une partie (ou une section) d'un corpus est comparée au corpus entier dans le but d'identifier le vocabulaire spécifique de la section. La comparaison partie-tout permet ainsi de décider si la fréquence relative d'un mot dans la section est normale ou non, et sinon, si elle est supérieure à ce qu'on pourrait prévoir en fonction de la fréquence relative du mot dans le corpus entier qui sert de point de référence. Cette méthode vise donc à mesurer les variations de fréquence dans un corpus découpé en parties (Labbé & Labbé 2001) et convient dès lors parfaitement à l'analyse d'un document constitué de plusieurs chapitres, pourvu qu'ils soient de longueur comparable. L'analyse statistique sous-jacente au calcul des spécificités utilise le test statistique de Fisher Exact[71], basé sur les probabilités exactes de la distribution hypergéométrique.

4.1.1.1 Pourquoi la distribution hypergéométrique ?

La distribution hypergéométrique est une distribution discrète des probabilités de valeurs, mais aussi des probabilités de fréquences de mots, par exemple. La distribution hypergéométrique décrit le nombre de réussites d'une séquence de n (nombre fixe) tirages exhaustifs, donc sans remise, dans une population finie. Les caractéristiques les plus importantes de la distribution hypergéométrique sont le fait qu'il n'y a pas de remise et que la population est discrète et finie (par exemple un corpus où le nombre de mots est discret et fini ou fixe).

L'exemple type permettant d'expliquer la distribution hypergéométrique est le tirage de boules rouges d'une urne contenant N boules, dont m boules rouges et $(N-m)$ boules blanches. On tire n boules de cette urne, sans remise, donc les boules tirées sont identifiées et mises de côté. Quelle est la probabilité (ou la chance) d'avoir tiré exactement k boules rouges ? La distribution des boules rouges dans l'échantillon des n boules tirées suit une distribution hypergéométrique, décrite par la formule générale (Cf. figure 4.1). A titre d'exemple, pour un échantillon ou un tirage sans remise de 8 boules d'une urne de 20 boules au total (dont 14 rouges et 6 blanches), on peut se demander quelle est la probabilité d'avoir exactement 6 boules

---

[70] Réimpression de l'édition de 1977. Müller est le fondateur de la statistique lexicale.

[71] Le test statistique de Fisher Exact est généralement utilisé pour des données de taille modeste, des corpus peu volumineux et des fréquences plutôt faibles ($n < 20$).

rouges[72] dans l'échantillon des 8 boules tirées. Il est clair que la probabilité d'être rouge change pour chaque boule qui est tirée, non seulement parce que le nombre total de boules change, mais également parce que la distribution des boules rouges et des boules blanches dans la population (l'urne) est modifiée.

$$\text{Prob } (X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

Figure 4.1 Formule générale de la distribution hypergéométrique

- *Intérêt de la distribution hypergéométrique pour un corpus linguistique*

On peut se demander quel est l'intérêt de la distribution hypergéométrique pour l'analyse d'un corpus linguistique ou pour l'analyse de données textuelles. En fait, une section ou une partie d'un corpus linguistique pourrait aussi être considérée comme un tirage exhaustif de mots (section) dans une population de mots (corpus entier). Chaque section est ainsi considérée comme un échantillon, permettant de situer la section « dans l'ensemble de tous les échantillons de même longueur qui peuvent être construits à partir du corpus » (Lafon 1984 : 54).

Pendant la délimitation d'une section, on n'inclut pas deux fois le même paragraphe, donc il n'y a pas de remise. Toutefois, le critère d'analyse pertinent n'est pas la couleur, mais la forme graphique des mots et le nombre total de mots dans le corpus (c'est-à-dire la taille du corpus). Après avoir pris le premier mot, le nombre total de mots compris dans le corpus change, et, par voie de conséquence, la fréquence relative des autres mots dans le corpus change aussi. La distribution hypergéométrique s'avère donc très appropriée pour l'analyse d'une section dans un

---

[72] Prob (X=6) = $\frac{\binom{14}{6}\binom{20-14}{8-6}}{\binom{20}{8}}$, où le coefficient binomial $\binom{20}{8}$ par exemple, indique le nombre de combinaisons possibles de 8 boules parmi 20 = $\frac{20!}{(20-8)!8!}$ ou 125.970 possibilités. A titre d'information, la probabilité Prob (X=6) est de 0,3576 (ou de 35,76%).

corpus linguistique ou dans un corpus de sections équivalentes. En effet, la distribution hypergéométrique permet de déterminer, non pas la probabilité d'avoir exactement 6 boules rouges, mais la probabilité d'avoir exactement la même fréquence relative d'un mot dans la section que dans la population (le corpus entier) ou d'avoir une fréquence relative déviante. La distribution hypergéométrique permet donc de calculer la probabilité que la fréquence observée d'un mot (dans la section) soit égale ou supérieure à la fréquence attendue de ce mot (la fréquence virtuelle basée sur la fréquence absolue dans le corpus entier).

Par conséquent, cette première approche méthodologique du calcul des spécificités consiste à déterminer si la fréquence d'un mot dans une section est normale (la fréquence observée égale la fréquence attendue, ou l'écart entre les deux est limité) ou si en revanche sa fréquence n'est pas normale (la fréquence observée est supérieure ou inférieure à la fréquence attendue). Au cas où la fréquence observée serait largement supérieure à la fréquence attendue, la probabilité calculée est très limitée.

La position méthodologique adoptée par Lafon (1984) est celle d'une comparaison partie-tout, prenant le tout comme étalon ou comme point de référence pour évaluer la partie. La formule de Lafon (1984 : 57) (Cf. figure 4.2) calcule la probabilité pour qu'un mot de fréquence f dans le corpus entier (de longueur T) apparaisse k fois dans la section i (de longueur t_i), dans l'hypothèse de l'équiprobabilité des sections.

$$\text{Prob } (X=k) = \frac{\binom{f}{k}\binom{T-f}{t_i-k}}{\binom{T}{t_i}}$$

Figure 4.2 Formule de la distribution hypergéométrique : corpus linguistique

La variable X suit une distribution hypergéométrique avec les paramètres T, t_i et f, et avec les contraintes suivantes sur k : « $f < t_i$ et $t_i < T - t_i$, et dans ce cas : $0 \leq k \leq f$ »[73] (Lafon 1984 : 57). Autrement dit, la fréquence f d'un mot dans le corpus entier doit être inférieure à la longueur de la section t_i, qui doit être inférieure à la longueur de

[73] Dans l'exemple concret de la probabilité des 6 boules rouges (si l'on tire 8 boules d'une urne de 20 boules), la contrainte $0 \leq k \leq m$ (f étant m) ou $0 \leq 6 \leq 14$ est respectée. Elle est compatible également avec la contrainte posée par Ross, à savoir $n-(N-m) \leq k \leq min(n, m)$ (Ross 1994), parce que $8-(20-14) \leq 6 \leq \min(8,14)$ ou $2 \leq 6 \leq \min(8,14)$.

l'ensemble des autres sections (T- t_i). Dans ce cas, la fréquence observée dans la section k doit être inférieure ou égale à f, ce qui est d'ailleurs toujours le cas pour une comparaison partie-tout.

- *Calcul de la probabilité dans un corpus linguistique*

La formule de la distribution hypergéométrique (Cf. figure 4.2), sous sa forme développée, se présente sous forme de factorielles et de produits[74]. En raison des factorielles, il est évident que « la taille des nombres obtenus atteint de telles dimensions qu'ils ne sont plus logeables en machine » (Lafon 1984 : 64-65). Dans un corpus de quelques milliers, voire de plusieurs millions de mots, les factorielles de la formule mèneraient à des nombres astronomiques[75]. Par conséquent, pour des nombres élevés tels que des fréquences dans un corpus linguistique, Lafon suggère de recourir à des logarithmes pour faciliter le calcul de la probabilité. Il propose dès lors la formule suivante (Lafon 1984 : 66) :

$$\log \text{Prob}\,(X=k) = \log f! + \log (T-f)! + \log t_i! + \log (T-t_i)! - \log T! - \log k! - \log (f-k)! - \log (t_i-k)! - \log (T-f-t_i+k)!$$

Figure 4.3 Formule du calcul de la probabilité dans un corpus linguistique

Quatre paramètres sont susceptibles de varier, à savoir la fréquence totale f du mot dans le corpus entier, sa fréquence dans la partie k, la taille du corpus T et la taille de la partie t_i (Labbé & Labbé 2001). Le résultat de ce nouveau calcul (Cf. figure 4.3) n'est pas la probabilité, mais le log de la probabilité. Par conséquent, un résultat de calcul tel que log Prob(X=k) = y, est à interpréter comme l'exposant de la base 10, d'où résulte une probabilité de 10^y.

[74] Le coefficient binomial $\binom{f}{k}$ s'écrit $\frac{f!}{(f-k)!k!}$. La factorielle de f, représentée par $f!$, est le produit des nombres entiers de 1 à f, sans omission ni répétition, à savoir $1\times2\times3\times\ldots\times f$. Il est à noter que la factorielle d'un nombre mène très vite à des nombres astronomiques, par exemple 10! = 1×2×3×4×5×6×7×8×9×10 = 3.628.800.

[75] « Les factorielles de la formule ne peuvent être programmées directement pour les chiffres auxquels on est confronté dans les corpus linguistiques (elles aboutissent à des nombres extraordinairement grands) » (Labbé & Labbé 2001).

A titre d'exemple, pour les boules rouges et blanches[76], la probabilité d'avoir exactement 6 boules rouges parmi 8 boules tirées, équivaut à $10^{-0,4466}$ ou à 0,3576.

Notons que dans un vaste corpus linguistique, il est également possible de recourir à des approximations de la distribution hypergéométrique pour calculer la probabilité des fréquences élevées, telles que des approximations binomiale, poissonienne et normale[77].

4.1.1.2 Résultats du calcul des spécificités : S^+ et S^-

De ce qui précède, il ressort que le résultat du calcul de la distribution hypergéométrique est une valeur de probabilité, qui indique la probabilité de la fréquence observée d'un mot dans une section, par rapport à sa fréquence dans le corpus entier, compte tenu de la taille des deux corpus. La valeur de probabilité indique donc si la fréquence dans la section est normale ou pas, par rapport au corpus entier. Après le calcul de cette probabilité, deux questions méthodologiques se posent. On peut se demander comment interpréter la valeur de probabilité pour pouvoir identifier les spécificités et comment passer de la valeur de probabilité au degré de spécificité.

Pour une section d'un corpus linguistique, découpé en plusieurs sections, la probabilité Prob ($X=k$) atteint un maximum à l'espérance mathématique (Labbé & Labbé 2001), c'est-à-dire lorsque la fréquence observée dans la section est égale à la fréquence attendue pour cette section, calculée à partir de la fréquence dans le corpus entier. Dans ce cas de figure, le mot apparaît aussi souvent qu'attendu et ce n'est pas une spécificité. Toutefois, si la fréquence observée est supérieure à la

---

76 Appliquons la formule avec les logarithmes à l'exemple de la probabilité des 6 boules rouges : log Prob(X=6) = … = log14!+log12!-log20!-log2!-log4! = -0,4466. Inscrit comme exposant de la base 10, le résultat du calcul permet d'obtenir la probabilité $10^{-0,4466} = 0,3576$.

77 L'approximation binomiale s'applique lorsque la longueur du corpus T et la fréquence totale du mot f sont très élevées par rapport à la taille de la section t_i, étant donné que la distribution binomiale (discrète) se caractérise par la remise, ce qui veut dire que les chances de réussite sont toujours égales (f/T). A son tour, la loi binomiale peut être approchée par la loi de Poisson (ou même par la loi normale), si t et f sont suffisamment grands.

L'approximation normale convient lorsque la section est infiniment large ($t \rightarrow \infty$) et elle se caractérise par un taux de réussite $p= ½$, pour $p=f/T$ (fréquence relative ou fréquence totale divisée par la taille du corpus entier). Malheureusement, l'identification de spécificités dans une section ne répond pas à ces critères. D'une part, la section n'est pas infiniment large et, d'autre part, les fréquences des mots dans un corpus sont très hétérogènes, donc le taux de réussite p serait différent pour chaque mot. Par conséquent, la distribution normale n'est pas appropriée pour les corpus linguistiques.

fréquence attendue, on calcule la probabilité $S^+ = \text{Prob}\ (X \geq k)$[78]. Si la probabilité est inférieure à un seuil défini (p < 0,05 ou p < 0,01), le mot sera qualifié de « spécificité positive »[79] (Lafon 1984) : il apparaît significativement plus souvent dans la section si on le compare à sa fréquence dans le corpus entier. Par contre, si la fréquence observée est inférieure à la fréquence attendue et si la probabilité $S^- = \text{Prob}\ (X \leq k)$[80] est inférieure à un seuil défini (p < 0,05 ou p < 0,01), le mot sera qualifié de « spécificité négative » (Lafon 1984), c'est-à-dire qu'il apparaît significativement moins souvent dans la section.

Ce sont surtout les spécificités positives qui sont intéressantes, car elles sont représentatives de la section, qu'elles caractérisent thématiquement. Notons qu'on pourrait aussi les reconnaître en lisant la section, car les spécificités y sont très fréquentes. Les spécificités négatives, en revanche, ne se laissent pas appréhender à travers une lecture simple de la section, car ces mots n'y figurent pas ou seulement très rarement. Pour l'identification des spécificités négatives, la confrontation avec le corpus entier est indispensable. Selon Lafon, « le relevé des spécificités négatives et positives d'une partie revient à *lire* à travers la *lunette* que constitue le corpus entier » (Lafon 1984 : 60).

La question se pose donc de savoir comment identifier automatiquement les spécificités et comment déterminer leur degré de spécificité. Le calcul des spécificités, qui est basé sur la distribution hypergéométrique et qui permet de déterminer le degré de spécificité, est implémenté dans les logiciels Lexico3[81] et

---

[78] Si la fréquence observée d'un mot dans une section est égale à 11, on calcule la probabilité que le mot apparaisse *au moins* 11 fois dans cette section, donc par exemple 11 fois ou 12 fois ou 13 fois ou même plus souvent. On calcule donc $\text{Prob}\ (X \geq 11)$ en additionnant $\text{Prob}(X=11) + \text{Prob}(X=12) + \text{Prob}(X=13) + \ldots$ (Lafon 1984).

[79] Il est à noter que le signe positif (+) ou négatif (-) du S ne fait pas partie du calcul hypergéométrique. Il est rajouté après le calcul pour faire la distinction entre les spécificités « positives » et « négatives ».

[80] Si la fréquence observée d'un mot dans une section est égale à 2, on calcule la probabilité que le mot apparaisse 2 fois *ou moins* dans cette section, donc 2 fois ou 1 fois ou 0 fois. On calcule $\text{Prob}\ (X \leq 2)$ en additionnant $\text{Prob}(X=2) + \text{Prob}(X=1) + \text{Prob}(X=0)$ (Lafon 1984).

[81] Lexico3 : SYLED – CLA2T, Paris3 :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/.

Hyperbase[82]. Dans Lexico3, ce sont les exposants (i.e. les résultats de la formule du calcul hypergéométrique avec les logarithmes) qui figurent dans la colonne du coefficient de spécificité, et non pas la probabilité elle-même. Plus élevé est le coefficient, plus faible sera la probabilité de la fréquence observée (par rapport au corpus entier) et plus spécifique sera le mot. Les spécificités positives indiquent un suremploi dans la section analysée, tandis que les spécificités négatives signalent un sous-emploi. Le coefficient de spécificité peut donc s'interpréter comme le degré de spécificité et permet dès lors de situer les spécificités sur une échelle ou un continuum de spécificité (Cf. annexe 6 pour l'utilisation pratique de Lexico3).

4.1.1.3 Recherches récentes

Le calcul des spécificités est surtout utilisé par la communauté francophone, notamment par Zimina (2004), Poibeau (2004) et Drouin (2003a et 2004). Zimina (2004) se sert du logiciel Lexico3 et du calcul des spécificités pour relever des spécificités dans des corpus alignés (français – anglais) d'environ 300.000 mots. Dans la pré-analyse de son corpus, Poibeau (2004) recourt à Lexico3 afin de caractériser a priori le corpus pour une application d'extraction d'information, parce que cet outil permet de combiner l'analyse des spécificités et la recherche des segments répétés.

Les recherches récentes de Drouin (2003a et 2004) recourent à la méthodologie du calcul des spécificités dans le but d'identifier des unités terminologiques (Lemay, L'Homme & Drouin 2005). Drouin (2004) compare un corpus d'analyse de langue spécialisée (12.000, 28.600 et 8.600 mots) à un corpus de référence de textes journalistiques (7,4 millions), portant sur des sujets variés. Cependant, il n'emploie pas le logiciel Lexico3 pour le calcul des spécificités (termes techniques), mais son propre logiciel TermoStat, un outil qui réunit le corpus d'analyse du domaine des télécommunications et le corpus de référence en un grand corpus hétérogène. Cette fusion des deux corpus est plutôt inhabituelle si on veut détecter des spécificités techniques dans un corpus technique par le biais d'un corpus de référence non technique. Le logiciel TermoStat implémente le calcul des spécificités par « une approximation normale de la loi hypergéométrique », telle que décrite par Lebart et

---

[82] Hyperbase : http://ancilla.unice.fr/default.html. Il convient de signaler que pendant le calcul des spécificités, Hyperbase prévoit tant la distribution hypergéométrique que la distribution normale. La loi hypergéométrique s'applique à tous les cas de figure. Toutefois, les résultats du calcul hypergéométrique se présentent sous forme de probabilités, souvent très faibles et toujours positives. « On doit avoir recours au signe de l'écart et à la représentation logarithmique. Hyperbase fait automatiquement le choix, plus exigeant, du modèle hypergéométrique lorsque la sécurité le recommande et que le corpus est de dimension restreinte » (Brunet 2002 : 36).

Salem (1994 : 182). Drouin (2004) respecte le seuil de 3,09 pour la valeur-test[83], ce qui lui permet de relever uniquement les mots qui ont moins d'une chance sur 1000 ($p < 0,001$) d'apparaître dans le corpus d'analyse avec la même fréquence relative ou avec une fréquence relative supérieure à celle du corpus de référence.

4.1.2 La méthode des mots-clés

La deuxième approche méthodologique pour l'identification des mots les plus spécifiques d'un corpus est couramment appelée *Keywords Method* ou méthode des mots-clés. Elle vise à comparer les fréquences dans un corpus de langue spécialisée aux fréquences dans un corpus de référence de langue générale, compte tenu de la taille des deux corpus, dans le but d'identifier les mots significativement plus fréquents dans le corpus spécialisé. Il s'agit donc de la comparaison de deux corpus différents, et non d'une comparaison *partie-tout*, caractéristique de la méthode du calcul des spécificités (Cf. 4.1.1). La deuxième différence entre les deux approches méthodologiques réside dans la statistique sous-jacente. La méthode des mots-clés se sert du rapport de vraisemblance (*log-likelihood ratio*) (Dunning1993). Cette statistique n'est pas basée sur des probabilités exactes, telles que les probabilités exactes de la distribution hypergéométrique, et par conséquent, elle s'applique plus facilement à des corpus plutôt volumineux. Le rapport de vraisemblance est surtout utilisé pour l'identification de mots spécifiques d'un domaine (mots-clés ou spécificités) et pour la détection de cooccurrences significatives (*composite terms*) (Dunning 1993). Il est à noter que la détection des cooccurrences significatives fera l'objet du deuxième axe méthodologique (Cf. chapitre 5). Dans cette section, nous expliquerons la mesure du rapport de vraisemblance[84] (LLR), ainsi que son importance pour la méthode des mots-clés et donc pour la détermination des spécificités dans un corpus spécialisé.

4.1.2.1 Pourquoi la mesure du rapport de vraisemblance ?

La méthode des mots-clés compare les fréquences relatives dans un corpus spécialisé aux fréquences relatives dans un corpus de référence, que l'on peut facilement représenter dans une table de contingence (Cf. tableau 4.1). La fréquence relative d'un mot dans un corpus exprime le rapport entre la fréquence absolue du mot et la taille du corpus, par exemple a/N_1 ou k/t pour le corpus spécialisé et b/N_2 pour le corpus de référence (en reprenant les codes k, t, f et T de la section 4.1.1).

---

[83] Un seuil de 3,09 correspond à une valeur $p < 0,001$.

[84] La mesure du rapport de vraisemblance (ou la mesure du LLR) est égale à $-2 \times \log$ du rapport de vraisemblance *(-2 log likelihood ratio*) (Cf. ci-dessous), mais on dit communément « (mesure du) rapport de vraisemblance » pour la désigner.

Afin de faciliter la comparaison des deux méthodes, nous considérons un grand corpus virtuel, comprenant le corpus spécialisé et le corpus de référence. Pour la fréquence *a* (ou *k*) dans le corpus spécialisé et la fréquence *b* dans le corpus de référence, *f* est la fréquence dans le corpus virtuel.

	Corpus spécialisé	Corpus de référence	Total = Corpus virtuel
Fréquence absolue	$\mathbf{a}$ $(= k)$	$\mathbf{b}$	$a + b \;= f$
Taille du corpus	$\mathbf{N_1}$ $(= t)$	$\mathbf{N_2}$	$N_1{+}N_2 = T$

Tableau 4.1 Table de contingence pour les fréquences relatives

Contrairement aux fréquences absolues, les fréquences relatives se prêtent bien à la comparaison de deux corpus. En plus, la fréquence relative d'un mot dans le corpus de référence pourrait également être considérée comme la fréquence attendue de ce mot dans le corpus spécialisé. Si la fréquence observée (fréquence relative dans le corpus spécialisé) est égale à la fréquence attendue (fréquence relative dans le corpus de référence), le mot apparaît aussi souvent que prévu et ce n'est pas un mot-clé spécifique du corpus spécialisé. Par contre, si la fréquence observée dépasse la fréquence attendue de façon statistiquement significative, le mot en question est un mot-clé spécifique du corpus spécialisé. Afin de déterminer si la différence de fréquence dans les deux corpus est statistiquement significative, le recours à une mesure statistique est indispensable.

- *Défauts et lacunes des autres mesures*

Afin de comparer les données dans une table de contingence, qui fait intervenir par exemple deux corpus linguistiques, plusieurs mesures statistiques sont disponibles, telles que le test du chi-carré (χ^2) de Pearson, le score Z et l'information mutuelle (*Mutual information* ou MI). Toutefois, les estimations de l'information mutuelle, qui sont basées directement sur les fréquences, ont tendance à surestimer la significativité des mots rares, c'est-à-dire des mots de faible fréquence. Par ailleurs, le score Z surestime considérablement la significativité des mots rares[85]. Et enfin, les valeurs du test du chi-carré (χ^2) de Pearson ne sont pas fiables pour des fréquences attendues inférieures à 5 ou même à 10 (Müller 1992a ; Rayson & Garside 2000). Cela s'explique principalement par le fait que l'hypothèse sous-jacente au score Z

---

[85] Pour les mots rares, les mesures statistiques, telles que le score Z, produisent des résultats peu fiables du point de vue statistique. Cependant, ces résultats sont parfois très utiles du point de vue terminologique (Lemay, L'Homme & Drouin 2005).

ainsi qu'au test du chi-carré est celle de la distribution normale[86], qui suppose que les mots analysés sont relativement fréquents (Dunning 1993).

Pour l'analyse de mots peu fréquents, la distribution normale ne convient pas, parce que les textes (ou les corpus textuels) contiennent en grande partie des mots peu fréquents. En effet, la répartition des fréquences dans un corpus ne suit pas de distribution normale, représentée traditionnellement par une courbe en cloche. Une distribution normale des fréquences indique qu'il y a peu de mots peu fréquents, peu de mots très fréquents, et beaucoup de mots moyennement fréquents. Cependant, dans les corpus linguistiques, il y a peu de mots très fréquents, un peu plus de mots moyennement fréquents, de plus en plus de mots qui sont de moins en moins fréquents et donc énormément de mots qui sont très peu fréquents ou même des hapax. La plupart des mots lexicaux (mots pleins) et des mots spécialisés sont en effet des mots (très) peu fréquents. Comme nous l'avons évoqué ci-dessus, la distribution normale n'est pas appropriée pour l'analyse de corpus linguistiques consistant en un majorité de mots rares, à moins que les corpus soient très vastes et que les analyses se limitent à des mots très fréquents.

- *La mesure du rapport de vraisemblance (log-likelihood ratio)*

Pour remédier au problème de l'analyse des mots peu fréquents dans les corpus linguistiques, Dunning (1993) propose la mesure statistique du rapport de vraisemblance (*Log-Likelihood Ratio* ou LLR ou encore G²). Celle-ci s'est avérée efficace tant pour des corpus vastes que pour des corpus restreints. Elle permet aussi la comparaison directe de la significativité de mots plus fréquents et de mots moins fréquents (Dunning 1993) en raison de son meilleur comportement asymptotique (approximatif). Par conséquent, la significativité des mots rares est plus fiable.

Tout comme le χ^2, le test du LLR est utilisé pour tester l'indépendance de deux variables multinomiales[87], en l'occurrence *k* et *f*, c'est-à-dire les fréquences observées dans le corpus spécialisé et dans le corpus de référence. Les valeurs des deux corpus (fréquence absolue et taille) sont considérées comme deux échantillons différents dont on veut savoir s'ils ont été prélevés dans la même population

---

[86] Quand la fréquence attendue (théorique) est « peu supérieure ou même inférieure à 1, il est peu indiqué de recourir au $\chi 2$, on préférera appliquer la loi de Poisson » (Müller 1992b : 53). La loi de Poisson convient par exemple lorsqu'on se réfère à un corpus de référence très étendu ou à un corpus de référence qui ne contient pas le texte (ou le corpus) analysé.

[87] Une variable binomiale peut avoir deux valeurs (p.ex. vrai ou faux), une variable multinomiale peut avoir plus de deux valeurs, par exemple la fréquence dans un corpus.

(Speelman 2005)[88]. Sous l'hypothèse nulle[89] (pas de différence de distribution ou pas de différence de fréquence relative dans les deux échantillons), les deux valeurs *k* et *f* relèvent de la même population et se caractérisent par la même distribution de fréquence (i.e. fréquence relative). Par contre, si l'hypothèse nulle est rejetée, les deux valeurs *k* et *f* relèvent de populations différentes et sont significativement différentes. Le test du LLR suppose que la distribution qui sous-tend les deux échantillons est une distribution binomiale[90]. Par conséquent, l'hypothèse nulle suppose que les fréquences absolues (ou observées) des deux corpus se caractérisent par la même probabilité sous-jacente de succès *p,* le succès étant la fréquence observée (Cf. table de contingence). Selon l'hypothèse alternative, les probabilités sous-jacentes de succès sont différentes dans les deux échantillons, en l'occurrence dans les deux corpus.

Ensuite, il faut déterminer les estimations d'échantillon pour les probabilités sous-jacentes dans les deux échantillons. A cette fin, le test du LLR recourt à la méthode de la vraisemblance ou de la probabilité maximale et, plus particulièrement, aux estimateurs du maximum de vraisemblance (ou vraisemblance maximale). Ces estimateurs sont les valeurs qui maximisent la probabilité de rencontrer ou d'observer exactement l'échantillon réalisé en question (i.e. chacune des deux colonnes de la table de contingence). Cela signifie que les valeurs de ces estimateurs seront déterminées de façon à maximiser la fonction qui exprime la probabilité d'observer les données de l'échantillon (fonction de probabilité ou *Likelihood function*). Les estimateurs sont inconnus et devront donc être estimés. Toutes les valeurs possibles de la probabilité de succès dans l'échantillon sont parcourues et pour chaque valeur de probabilité de succès, la probabilité d'avoir cet échantillon est

---

88 Pour des explications plus détaillées : Cf. *chapter 4 « Words and the company they keep »* (section : mots-clés et collocations) du cours « *Methods of Corpus Linguistics* ».

89 « Pour démontrer la validité d'une hypothèse, la démarche statistique consiste en général à lui opposer une hypothèse nulle, et à décider d'un intervalle à l'intérieur duquel il serait imprudent de rejeter l'hypothèse nulle, donc d'adopter l'hypothèse contraire » (Müller 1992a : 91).

90 Une distribution binomiale est une distribution de probabilité discrète, caractérisée par deux résultats possibles : succès ou échec, par exemple lancer une pièce de monnaie, où la probabilité de réussite est $p = ½$. Le fait de compter le nombre de succès d'événements répétés identiques et indépendants, se prête bien au dénombrement de mots dans un texte ou corpus, c'est-à-dire aux fréquences de mots. Chaque occurrence dans le texte est comparée au mot qui est compté, ce qui permet de compter les succès du mot en question, c'est-à-dire la fréquence absolue (observée) de ce mot. Toutefois, il est à noter que dans un texte ou corpus, les mots n'apparaissent pas tout à fait indépendamment les uns des autres. Mais, au fur et à mesure de la distance, la dépendance des mots diminue (Dunning 1993).

déterminée. La valeur pour laquelle la probabilité d'avoir cet échantillon est maximale sera l'estimateur du maximum de vraisemblance.

D'après la distribution binomiale, on sait que la chance ou la probabilité de succès (pour la fréquence observée de la table de contingence) correspond à la fréquence relative. Cependant, ce n'est pas la valeur de probabilité de succès (ou l'estimateur du maximum de vraisemblance) qui est requise pour le test du rapport de vraisemblance, mais la probabilité maximale correspondante, à savoir la probabilité maximale (calculée) d'avoir cet échantillon, étant donné l'estimateur.

Le rapport de vraisemblance ou le rapport de probabilité exprime le rapport entre deux probabilités maximales : dans le numérateur (1) la probabilité maximale sous l'hypothèse nulle que les deux échantillons (deux colonnes) relèvent de la même population et dans le dénominateur (2) la probabilité maximale en général. Notons qu'il s'agit de la probabilité maximale combinée pour les deux échantillons (i.e. les deux colonnes de la table de contingence).

Le numérateur du rapport de vraisemblance (1) est la probabilité maximale combinée qui prévoit d'avoir en même temps les valeurs des deux colonnes de la table de contingence (donc les deux fréquences observées k et f), sous l'hypothèse nulle de la même probabilité sous-jacente. Les deux estimateurs du maximum de vraisemblance devront donc être identiques. On considère ensuite la probabilité maximale combinée des deux échantillons, étant donné que ces estimateurs sont identiques. Le dénominateur de ce rapport (2) est la probabilité maximale combinée des deux fréquences observées séparément, autrement dit les estimateurs de probabilité ne doivent pas être identiques. Le dénominateur donne la probabilité maximale en général, dans des circonstances optimales. Il est clair que, pour un mot spécifique du corpus spécialisé, la probabilité maximale combinée sous l'hypothèse nulle sera très faible, parce que les fréquences observées (dans les deux corpus) n'auront quasiment jamais la même probabilité sous-jacente.

Le rapport de vraisemblance (ou le rapport de probabilité) se situe entre 0 et 1. Plus ce rapport s'approche de 0, plus la probabilité maximale du numérateur (sous l'hypothèse nulle) sera faible, donc plus l'hypothèse nulle sera improbable et plus l'écart par rapport à l'hypothèse nulle de non-différence sera grand. Toutefois, la mesure statistique du rapport de vraisemblance G^2 (ou LLR) (*log-likelihood ratio*) n'est pas égale au rapport de vraisemblance calculé ci-dessus (soit V), mais au log de ce rapport multiplié par -2, donc $-2 \times \log(V)$. Cette transformation permet d'obtenir une quantité ($-2 \log \lambda$) qui suit une distribution connue, en l'occurrence une

distribution χ^2 asymptotiquement[91]. On peut dès lors déterminer la valeur de probabilité associée au résultat du calcul. Lorsque le résultat G² ou LLR est supérieur ou égal à 3,84, on peut rejeter l'hypothèse nulle (pas de différence significative) avec une confiance de 95%.

Après les deux opérations mathématiques (le logarithme et la multiplication), des valeurs élevées pour la mesure statistique G² ou LLR indiquent un écart plus important de l'hypothèse nulle. Du point de vue de l'opposition entre le corpus spécialisé et le corpus de référence, une valeur plus élevée pour la mesure statistique G² ou LLR signifie que le mot en question sera plus spécifique dans le corpus spécialisé, par rapport au corpus de référence de langue générale. Dès lors, ce mot est identifié comme étant un mot spécifique ou un mot-clé du corpus spécialisé. Comme nous l'avons évoqué ci-dessus, les mots-clés ou les spécificités se caractérisent par une valeur très élevée de LLR et par une valeur de probabilité (valeur p) associée très faible.

- *Calcul du rapport de vraisemblance dans des corpus linguistiques*

Pour le calcul du rapport de vraisemblance (G² ou LLR), on aura besoin de la taille des deux corpus (corpus spécialisé et corpus de référence) et de la fréquence absolue d'un mot dans les deux corpus, visualisées par N_1, N_2, a et b (Cf. tableau 4.2).

	Corpus spécialisé	Corpus de référence	Total
Fréquence du mot	$\boldsymbol{a}$	$\boldsymbol{b}$	$a + b$
Fréquence des autres mots	$\neg a$ $(N_1 - a)$	$\neg b$ $(N_2 - b)$	$\neg a + \neg b$
Taille du corpus	$\boldsymbol{N_1}$	$\boldsymbol{N_2}$	$N\ (N_1 + N_2)$

Tableau 4.2 Table de contingence pour la comparaison de fréquences

Les valeurs observées (*observed values*) sont $O_1 = a$ et $O_2 = b$. Les valeurs attendues E_1 et E_2 (*expected values*) sont calculées en fonction de la taille des deux corpus[92] : donc $E_1 = N_1*(a+b)/(N_1+N_2)$ et $E_2 = N_2*(a+b)/(N_1+N_2)$ (Rayson & Garside 2000 : 3). La prise en compte de la taille des deux corpus permet

---

91 Pour les détails mathématiques, nous renvoyons à Dunning (1993).

92 Selon la formule suivante : $E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$ (Rayson & Garside 2000 : 3).

d'appliquer tout de suite la formule[93] pour le calcul de la valeur du rapport de vraisemblance, ce qui revient à calculer le *log-likelihood ratio* comme suit (Rayson & Garside 2000 : 3) (Cf. figure 4.4)[94].

$$LLR = 2*((a*log(a/E_1)) + (b*log(b/E_2)))$$

Figure 4.4 Formule du calcul du rapport de vraisemblance

Des valeurs élevées de LLR (ou *log-likelihood ratio*) indiquent une différence très significative entre les fréquences relatives dans les deux corpus. Par conséquent, les mots avec les valeurs de LLR les plus élevées sont les plus spécifiques d'un des deux corpus. Les mots ayant des fréquences relatives comparables dans les deux corpus ne sont pas spécifiques.

4.1.2.2 Résultats de la méthode des mots-clés

Le résultat du calcul de la mesure statistique du rapport de vraisemblance est une valeur de spécificité (valeur de LLR), qui indique directement le degré de spécificité du mot, par le biais de la comparaison des fréquences relatives dans les deux corpus. Le rapport de vraisemblance (LLR) sera d'autant plus élevé que le mot est plus fréquent dans le corpus spécialisé par rapport au corpus de référence. Une variable supplémentaire dans le fichier de sortie des logiciels et des outils indiquera s'il s'agit d'une spécificité positive et donc d'un suremploi dans le corpus spécialisé ou s'il s'agit d'une spécificité négative et donc d'un sous-emploi dans le corpus spécialisé. La valeur p correspondante permet de supprimer les spécificités statistiquement non significatives ($p < 0,05$). Ce seuil de significativité correspond à une valeur de LLR supérieure à 3,84 environ (en fonction des corpus). La mesure statistique de test LLR est une mesure statistique solide, qui convient très bien à des corpus volumineux et qui permet la comparaison de la significativité des mots peu fréquents[95] et de ceux qui sont plus fréquents.

Notons que le tri des spécificités en fonction de la mesure statistique du LLR (rapport de vraisemblance) permet de classer les spécificités par ordre de spécificité

---

[93] Formule du G² ou LLR = $2 \sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$ (Rayson & Garside 2000 : 3).

[94] Un calculateur du rapport de vraisemblance (LLR) (pour les fréquences d'un mot dans deux corpus) est disponible sur : http://ucrel.lancs.ac.uk/llwizard.html.

[95] Il est généralement admis que la fréquence attendue devra être supérieure à 3.

décroissante et par conséquent, de les situer sur une échelle de spécificité ou un continuum de spécificité.

4.1.2.3 Recherches récentes

La méthode des mots-clés, qui permet d'identifier les mots les plus spécifiques d'un corpus spécialisé, est surtout utilisée par des utilisateurs du logiciel WordSmith, et plus particulièrement de l'outil KeyWords[96] (Cf. 4.2.3), et donc par la communauté anglophone. La méthode des mots-clés est également implémentée dans le logiciel Abundantia Verborum Frequency List Tool[97]. Citons notamment les travaux de Berber Sardinha (1996, 1999a et 1999b) et Vangehuchten (2004), qui ont eu recours à cette approche méthodologique.

Les travaux de Berber-Sardinha décrivent l'identification de mots-clés à l'aide du logiciel WordSmith KeyWords (1999a et 1999b). Un corpus d'analyse restreint de rapports d'affaires (3.355 occurrences) est comparé à un corpus de référence de 17 rapports (95.541 occurrences)[98] dans le but de déterminer les mots les plus spécifiques du corpus d'analyse. L'extraction des mots-clés et de leurs collocations s'inscrit dans le cadre d'une étude d'identification des thèmes d'un texte, étant donné que les mots spécifiques ou mots-clés reflètent le contenu thématique principal du texte analysé.

Les recherches de Vangehuchten (2004) se situent dans le contexte didactique de l'espagnol pour objectifs spécifiques. Afin de procéder à une sélection objective du vocabulaire à enseigner, un corpus d'analyse de textes spécialisés d'environ 120.000 occurrences (manuel espagnol d'introduction à l'économie d'entreprise) est comparé à un corpus de référence de langue générale de 19,4 millions d'occurrences. La mesure statistique du LLR permet de déterminer la sélection objective, car statistiquement significative, des mots spécifiques du corpus.

---

[96] WordSmith Tools version 3 : http://www.lexically.net/wordsmith/.

[97] Abundantia Verborum : http://wwwling.arts.kuleuven.be/genling/abundant/obtain/.

[98] Il est à noter que ce rapport ne correspond pas au rapport de 1/10.

4.2 ÉTUDE COMPARÉE DE TROIS OUTILS

Les recherches récentes en matière d'identification de spécificités recourent principalement aux outils Lexico3, WordSmith et Abundantia Verborum (Cf. 4.1.1 et 4.1.2). Le logiciel Lexico3 s'appuie sur la première approche méthodologique du calcul des spécificités et de la distribution hypergéométrique. Les deux autres logiciels, WordSmith et Abundantia Verborum, utilisent la méthode des mots-clés et sa statistique sous-jacente du rapport de vraisemblance (LLR). L'utilisation pratique de ces trois outils est expliquée dans le document en annexe (Cf. annexe 6). Dans cette partie, nous procédons à une comparaison quantitative et qualitative des résultats des deux approches méthodologiques et des trois outils principaux.

Dans leur étude comparée, Lemay, L'Homme et Drouin (2005) évaluent deux méthodes pour l'identification et l'extraction d'unités terminologiques unilexicales (unités simples) dans un corpus spécialisé. Les deux méthodes reposent sur la première approche méthodologique du calcul des spécificités, implémentée dans le logiciel TermoStat (Drouin 2003a et 2004). Toutefois, elles se distinguent par le corpus de référence. La première méthode compare un corpus d'analyse de langue spécialisée (600.000 mots) relevant du domaine de l'informatique, à un corpus de référence de langue générale (30 millions de mots) du journal *Le Monde*. Dans la deuxième méthode, chacun des six sous-corpus thématiques du corpus d'analyse spécialisé est comparé au corpus spécialisé entier, servant alors de corpus de référence, ce qui est similaire à l'approche partie-tout de Lafon (1984) et de Lebart et Salem (1994). Ces six listes de spécificités sont alors réunies en une seule liste de spécificités pour la deuxième méthode. Ensuite, on évalue les deux méthodes en comparant les deux listes de spécificités au contenu de deux dictionnaires spécialisés relevant du même domaine spécialisé. Les résultats de la première méthode sont légèrement meilleurs, en termes de précision et de rappel, mais les deux méthodes sont utiles pour l'extraction d'unités terminologiques unilexicales.

Afin de comparer les résultats des deux approches méthodologiques et des trois outils cités ci-dessus, nous avons conduit plusieurs expérimentations sur la version lemmatisée de deux corpus de test : un petit échantillon du corpus technique spécialisé (690 lignes de texte ou environ 14.000 occurrences (80 Ko)) et un échantillon du corpus de référence du journal *Le Monde* (6314 lignes ou environ 106.000 occurrences (635 Ko)). Les deux échantillons réunis constituent le corpus servant à ces expérimentations (7004 lignes de texte suivi lemmatisé ou environ 120.000 occurrences (715 Ko)). Lorsque le corpus spécialisé est comparé au corpus entier, le rapport partie-tout est de 1/10 approximativement. Les détails de ces expérimentations sont explicités en annexe (Cf. annexe 6).

Généralement, les trois outils cités ci-dessus conduisent à des résultats similaires en ce qui concerne les spécificités relevées, le nombre de spécificités et le rang des spécificités. Les similarités seront précisées dans la première section de cette deuxième partie (4.2.1). Dans la deuxième section (4.2.2), nous expliquerons les différences les plus importantes entre les trois outils, principalement en ce qui concerne la valeur de probabilité, le coefficient de spécificité et le corpus de référence. Les différences proviennent essentiellement de la méthodologie et de la statistique sous-jacentes, comme nous l'avons déjà évoqué ci-dessus (Cf. 4.1).

4.2.1 Similarités

Dans les trois outils, le calcul des spécificités ou l'identification des mots-clés s'effectue à un seuil de significativité de 0,05 ($p < 0,05$) et inclut tous les mots (au niveau des lemmes) du corpus de test. La comparaison globale du nombre de spécificités relevées (Cf. tableau 4.3) confirme le fait que Abundantia Verborum (désormais AV) et WordSmith (désormais WS) s'appuient tous les deux sur la même méthodologie de la méthode des mots-clés (*Keywords Method*), qui compare le corpus spécialisé à un corpus de référence. Lexico3, par contre, compare le corpus spécialisé au corpus entier.

	AV	WS	Lexico3
Spécificités positives	885	873	1666

Tableau 4.3 Nombre de spécificités positives dans les trois outils

Il est à noter que la liste de spécificités de Lexico3 de cette expérimentation est beaucoup plus longue que les deux autres listes de spécificités. La différence de nombre de spécificités s'explique notamment par le fait que Lexico3 inclut aussi des nombres parmi les spécificités, à la différence de AV et WS[99]. Toutefois, la raison principale réside dans le coefficient de spécificité ou l'indice de spécificité, qui indique indirectement la valeur de probabilité. Les spécificités en fin de liste (rangs

[99] Même si les données chiffrées et les nombres ne sont pas intéressants en tant que mots-clés pour l'analyse sémantique quantitative, il reste possible d'inclure des nombres dans les listes de mots-clés de AV et de WS. Dans AV, les listes de fréquence générées dans le logiciel AV proprement dit (version *wordlist*) recensent des données chiffrées et des nombres. Toutefois, ces derniers sont filtrés dans l'outil AV Frequency List Tool, au moment de générer la liste de mots-clés. La désactivation du filtrage des nombres aboutit à une liste de 1103 mots-clés. Dans WS, le filtrage des nombres s'effectue au moment de générer les listes de fréquence, c'est-à-dire avant de passer à l'outil Keywords. A titre de comparaison, cette liste de mots-clés dans WS comprend 1102 mots-clés.

de 1450 à 1666), c'est-à-dire les spécificités les moins spécifiques, ont toutes un coefficient de 1. Ce coefficient de spécificité équivaut à une probabilité de 10^{-1} ou de 0,1, qui n'est pas toujours statistiquement significative. Par conséquent, les mots à la fin de la liste de spécificités de Lexico3 ne sont pas tous des spécificités statistiquement significatives (au seuil de significativité statistiquement pertinent de 0,05), comme nous l'expliquerons dans la section suivante (Cf. 4.2.2). Globalement, les outils AV et WS génèrent des listes de spécificités comparables à la liste de spécificités de Lexico3, surtout si l'on compare les spécificités les plus pertinentes, qui figurent en tête de liste (Cf. tableau 4.4).

	AV	LLR	WS	KEYNESS	Lexico3	Fq. Tot.	Fq	Coeff.
1	*axe*	668,63	*AXE*	658,7	*mn*	98	98	50
2	*usinage*	590,32	*USINAGE*	581,5	*usinage*	136	136	50
3	*x*	552,59	*X*	537,5	*outil*	75	68	50
4	*mm*	435,03	*MM*	428,1	*pièce*	101	80	50
5	*mn*	425,13	*MN*	418,8	*jusque*	56	56	50
6	*broche*	260,13	*BROCHE*	256,3	*mm*	112	108	50
7	*outil*	250,02	*OUTIL*	245,7	*vitesse*	69	63	50
8	*pièce*	248,77	*PIÈCE*	243,8	*axe*	154	154	50
9	*jusque*	242,77	*JUSQUE*	239,2	*broche*	60	60	50
10	*vitesse*	233,83	*VITESSE*	229,8	*x*	142	137	50
11	*plaquette*	195,05	*PLAQUETTE*	192,2	*plaquette*	45	45	42
12	*un*	170,89	*KW*	162,3	-	97	68	40
13	*kw*	164,69	*FRAISE*	153,7	*kw*	38	38	36
14	*fraise*	156,02	*UN*	151,5	*fraise*	36	36	34
15	*table*	147,32	*TABLE*	144,7	*un*	3526	664	34
16	*à*	139,70	*ROTATIF*	136,6	*table*	48	42	33
17	*machine*	139,11	*MACHINE*	135,8	*machine*	82	54	30
18	*rotatif*	138,67	*DIAMÈTRE*	132,3	*rotatif*	32	32	30
19	*diamètre*	134,34	*À*	124,9	*centre*	98	58	29
20	*centre*	128,62	*Z*	121,9	*diamètre*	31	31	29
21	*z*	119,73	*CENTRE*	121	*à*	2504	490	29
22	*degré*	110,98	*DEGRÉ*	108,9	*z*	33	31	27
23	*course*	110,58	*COURSE*	108,6	*degré*	39	33	25
24	*nuance*	110,43	*NUANCE*	108,5	*cinq*	101	55	25
25	*cinq*	110,38	*CINQ*	107,2	*course*	37	32	25
26	*avance*	108,61	*AVANCE*	106,9	*nuance*	35	31	25
27	*usiner*	108,33	*USINER*	106,7	*usiner*	25	25	24
28	*acier*	103,99	*ACIER*	102,5	*avance*	28	27	24
29	*précision*	102,93	*PRÉCISION*	101,2	*précision*	29	27	23
30	*Trametal*	99,66	*TRAMETAL*	98,2	*acier*	24	24	23

Tableau 4.4 Résultats des trois outils : les 30 mots les plus spécifiques

4.2.2 Différences

Les principales différences entre les trois outils (et entre les deux approches méthodologiques) se situent à deux niveaux : d'une part, au niveau de la valeur de probabilité et du coefficient de spécificité et, d'autre part, au niveau du corpus de référence. La valeur de probabilité et le coefficient de spécificité sont toujours associés et relèvent de la mesure statistique sous-jacente. La différence en matière de corpus de référence réside dans le choix méthodologique (partie-tout versus spécialisé-général). Etant donné que ces différences affectent les spécificités relevées et leur degré de spécificité, elles joueront un rôle important dans le choix de l'approche méthodologique la plus appropriée.

- *Valeur de probabilité*

La valeur de probabilité ou la valeur p est associée à une mesure statistique telle que la mesure du LLR. La valeur de probabilité indique le seuil de rejet ou le seuil de significativité, permettant de savoir si l'hypothèse nulle d'indépendance (pas de différence significative) est vérifiée et donc si les différences observées sont dues au hasard. Sous l'hypothèse nulle, il n'y a pas de différence entre la fréquence observée et la fréquence attendue, c'est-à-dire entre les fréquences dans le corpus spécialisé et dans le corpus de référence. Il est généralement admis que l'hypothèse nulle peut être rejetée à un seuil de rejet ou à un seuil de significativité de 5% (valeur $p < 0,05$). A ce seuil, la probabilité de se tromper en rejetant l'hypothèse nulle d'indépendance est de 5%, ce qui veut dire que dans 5% des cas, on se trompe en rejetant l'hypothèse nulle d'indépendance. Par conséquent, les différences observées sont dites statistiquement significatives. La chance ou la probabilité qu'elles soient dues au hasard n'est que de 5%. Les seuils de significativité couramment utilisés sont les seuils de 0,05 et de 0,01 (ou même de 0,001). Des seuils plus bas sont plus convaincants, puisqu'ils sont plus fiables. Par contre, des résultats observés à un seuil de significativité de 0,05, pour une valeur p inférieure à 0,05 (marge d'erreur de 5%), se situent aux confins de la significativité, mais ils sont statistiquement significatifs. Au seuil de 0,01 ($p < 0,01$), les résultats observés sont communément considérés comme statistiquement significatifs (marge d'erreur de 1%) et au seuil de 0,001, ils sont hautement significatifs (marge d'erreur de 0,1%).

Dans AV, la valeur p par défaut est fixée à 0,05. L'exportation des résultats dans Excel permet de trier par ordre décroissant de spécificité (LLR) ou de valeur p (plus particulièrement en fonction du complément de la valeur p ou (1-p)). Couper en fonction d'une valeur $p < 0,001$ par exemple, permet d'être plus sévère et de ne retenir que les spécificités les plus pertinentes, les plus fiables et les plus significatives, qui sont moins nombreuses. A l'instar de l'outil AV, la valeur p dans WS est fixée à 0,05. Dans Lexico3, il y a 2 paramètres modifiables, à savoir la

fréquence minimale et la valeur p. La valeur $p < 0,01$ est la valeur de probabilité la plus faible possible dans Lexico3.

Il est à noter que dans les résultats du calcul des spécificités dans Lexico3, la valeur de probabilité est obtenue de façon indirecte, à partir du coefficient de spécificité, qui est toujours un nombre entier. Ainsi, un coefficient de spécificité de 2 indique une probabilité de 10^{-2} (ou $p < 0,01$) et un coefficient de spécificité de 3 signifie une probabilité de 10^{-3} (ou $p < 0,001$), etc. Le coefficient 1 indique une valeur de probabilité de 0,1, ce qui n'est pas toujours statistiquement significatif, même si la fenêtre de paramétrage prévoit un seuil de significativité équivalent à $p < 0,05$. Par conséquent, les spécificités en bas de liste (avec le coefficient 1) devront être exclues des analyses, pour que le seuil de significativité soit inférieur à 0,01 et donc statistiquement significatif et fiable.

Dans le cadre des expérimentations de l'étude comparée, différents paramètres ont été testés dans Lexico3 pour la valeur p ($p < 0,05$ et $p < 0,01$). Le même seuil de significativité dans les logiciels AV et WS permet de procéder à une comparaison détaillée des résultats des trois outils (Cf. tableau 4.5), qui correspond à une comparaison du nombre de spécificités positives en fonction de plusieurs seuils de significativité.

	Avec coeff. 1 ou p < 0,1	Sans coeff. 1 (0,1) ou p < 0,01	Sans coeff. 2 (0,01) ou p < 0,001
Lexico3prob5fq1	1666	660	393
AV	885	445	295
WS	873	433	286

Tableau 4.5 Nombre de spécificités positives dans les 3 outils pour 3 seuils

Il est clair que pour les spécificités les plus pertinentes et les plus fiables (donc aux seuils de $p < 0,01$ et $p < 0,001$), les résultats sont les plus convergents (Cf. les deux dernières colonnes du tableau 4.5). Toutefois, rappelons que les listes de spécificités de Lexico3 sont un peu plus longues (entre 2 et 1,5 fois plus longues), puisqu'elles comprennent également des données chiffrées et des nombres (Cf. ci-dessus).

- *Coefficient de spécificité*

La valeur de probabilité ou le seuil de significativité est associé à un coefficient de spécificité dans AV et WS. Ces outils permettent de trier les spécificités en fonction de la mesure statistique du LLR (*keyness* dans WS). Ainsi, les mots les plus spécifiques du corpus spécialisé figurent en tête de liste. Dès lors, la valeur de LLR indique le degré de spécificité et permet de situer les spécificités sur un continuum de spécificité, allant des plus spécifiques aux moins spécifiques, mais incluant

toujours des spécificités statistiquement significatives. Cependant, dans une liste de plusieurs milliers de spécificités, les mots dont la fréquence technique et la fréquence générale sont identiques se voient attribuer la même valeur de LLR et donc le même degré de spécificité. Or, cela ne pose pas de problème interprétatif, car d'un point de vue purement quantitatif, les mots avec la même fréquence technique et la même fréquence générale se situent au même niveau de spécificité et se verront par conséquent attribuer le même rang de spécificité (Cf. chapitre 6). En dépit du fait que les outils AV Frequency List Tool et WS Keywords reposent sur la même méthodologie et la même statistique sous-jacente (la mesure du LLR), ils génèrent tout de même des listes de mots-clés ou de spécificités légèrement différentes[100]. D'une part, on observe une différence en ce qui concerne l'orthographe des mots-clés. D'autre part, la valeur de LLR (*keyness* dans WS) présente quelques fluctuations lorsqu'on compare les deux listes (Cf. tableau 4.4).

Ces différences s'expliquent principalement par les caractéristiques des listes de fréquence de ces deux outils. Les listes de fréquence de WS sont générées dans l'outil WS Wordlist et maintiennent les mots avec trait d'union. Par contre, la fonctionnalité des listes de fréquence (version *wordlist*) du logiciel Abundantia Verborum, utilisée dans cette expérimentation, ne prend pas en considération le trait d'union ni l'apostrophe comme délimitateur de mots. Ainsi, des mots tels que *machine-outil* et *aujourd'hui* sont repris dans les listes de fréquence AV (version *wordlist*) comme *machineoutil* et *aujourdhui*, sans trait d'union ni apostrophe. En plus, même si les fréquences absolues sont quasi identiques dans les listes de fréquence des deux outils, la valeur de LLR présente quand même quelques fluctuations. En effet, la valeur de LLR n'est pas déterminée en fonction des fréquences absolues, mais en fonction des fréquences relatives. Autrement dit, on tient compte du nombre total d'occurrences dans le corpus spécialisé et dans le corpus général, à partir des données des listes de fréquence. Or, le nombre total d'occurrences dans les listes de fréquence de AV (version *wordlist*) est légèrement supérieur au nombre total d'occurrences dans les listes de fréquence générées dans WS (Cf. tableau 4.6).

Outil	Nombre total d'occurrences dans le corpus spécialisé (1)	Nombre total d'occurrences dans le corpus général (2)
AV (version *worlist*)	14.303	106.021
WS (outil Wordlist)	14.207	105.785

Tableau 4.6 Nombre total d'occurrences (listes de fréquence de AV et de WS)

[100] Le taux de recoupement se situe entre 97% et 98%.

Ces différences concernant le nombre total d'occurrences sont dues essentiellement aux particularités typographiques des listes des deux outils (p.ex. µ dans AV, mais pas dans WS). Par conséquent, le calcul du LLR, même s'il est basé sur la même formule, conduit à des résultats légèrement différents (Cf. tableau 4.4). Il s'ensuit que les spécificités avec trait d'union ou avec apostrophe figurent uniquement dans la liste de mots-clés de WS. Dans la liste de AV, on recense la variante de ces spécificités, mais en un mot. Les détails de ces différences entre les listes de spécificités de AV et de WS sont consignées en annexe (Cf. annexe 6 : tableaux A6.4 et A6.5).

Dans AV, il est possible de résoudre le problème de la version *wordlist* qui supprime le trait d'union et l'apostrophe. Pour cela, il faut générer les listes de fréquence à partir d'un script en Python, qui parcourt les fichiers *.cnr, c'est-à-dire les fichiers des deux corpus, produits par l'analyseur Cordial (Cf. chapitre 3). Le recours au script permet en effet plus de flexibilité au niveau des listes de fréquence : soit des listes de formes fléchies, soit des listes de lemmes, soit des listes en fonction du code de la classe lexicale. En plus, les listes de fréquence des lemmes s'avèrent très adéquates, puisqu'elles comprennent les lemmes, tels qu'ils sont catégorisés par Cordial. Ces nouvelles listes de fréquence recensent, par exemple, les lemmes *machine-outil* ou *aujourd'hui*, mais aussi les lemmes *afin de* ou *t/mn* (Cf. annexe 6 : tableaux A6.6 et A6.7).

Lexico3 n'affiche pas de mesure statistique, parce qu'il n'y en a pas. Le calcul de la distribution hypergéométrique exacte consiste à calculer la probabilité ou du moins le log de la probabilité. Du point de vue mathématique, le calcul mène à un résultat, à savoir le log de la probabilité, implémenté dans Lexico3 comme le coefficient de spécificité ou l'indice de spécificité. Ce coefficient de spécificité permet de trier les spécificités, car elles sont affichées par ordre décroissant de spécificité. Toutefois, il est à noter que le coefficient de spécificité de Lexico3 dans cette expérimentation est un nombre entier et arrondi entre 1 et 50. Celui-ci ne permet pas d'opérer des distinctions aussi fines que celles de la valeur de LLR (Cf. AV et WS). En plus, le même coefficient de spécificité est attribué à un nombre très important de spécificités, ce qui empêche un classement très précis. Ainsi, dans les expérimentations impliquant un échantillon du corpus spécialisé de 14.000 occurrences (version lemmatisée), les 10 mots les plus spécifiques ont tous le coefficient 50 (probabilité de 10^{-50}). Pour les coefficients plus bas, tels que 1 et 2 (probabilité de 10^{-2}), la granularité est encore plus problématique. Dans la liste de spécificités qui inclut tous les mots, le coefficient 1 caractérise pas moins de 1006 spécificités, alors que le coefficient 2 est attribué à 267 spécificités.

- *Corpus de référence*

Pour comparer le corpus spécialisé à un corpus de référence, deux possibilités sont à envisager : (1) la comparaison partie-tout, où le corpus de référence est le corpus entier et (2) la comparaison d'un corpus spécialisé à un corpus de référence, qui est un corpus de langue générale, indépendant du corpus spécialisé. Ces deux types de corpus de référence sont liés aux deux approches méthodologiques évoquées ci-dessus. Premièrement, la comparaison partie-tout, caractéristique du calcul des spécificités (Cf. 4.1.1), compare une section du corpus à l'ensemble du corpus. Par conséquent, en vue de son application à notre corpus spécialisé, cette approche méthodologique, le type de corpus de référence et principalement la statistique sous-jacente de la distribution hypergéométrique requièrent l'incorporation du corpus spécialisé dans le corpus de langue générale en vue de construire un grand corpus de référence virtuel. Du point de vue méthodologique, cette incorporation n'est pas très satisfaisante, parce que le corpus de référence devient hétérogène s'il comprend une section spécialisée et neuf sections générales. Le deuxième type de corpus de référence, caractéristique de la méthode des mots-clés (Cf. 4.1.2), compare le corpus spécialisé à un corpus de référence de langue générale indépendant qui n'inclut pas le corpus spécialisé.

Lexico3 procède par comparaison partie-tout (1), tandis que AV et WS comparent le corpus spécialisé à un corpus de référence de langue générale (2). Or, pour le grand corpus entier virtuel qui compte 17 millions d'occurrences (1,7 million + 15,3 millions), le traitement informatique[101] dans Lexico3 poserait des problèmes de vitesse de calcul et de mémoire vive, en raison des fréquences considérables et du nombre de mots à traiter.

On pourrait envisager, à titre d'expérimentation, de recourir également à la première modalité de corpus de référence (partie-tout) dans AV et WS, étant donné que Lexico3 ne permet pas d'autres types de comparaison de corpus. Toutefois, dans AV et WS, la comparaison du corpus spécialisé au corpus entier n'est pas correcte du point de vue méthodologique[102], ce qui se reflète d'ailleurs dans les résultats (Cf. tableau 4.7). En effet, la comparaison partie-tout (i.e. corpus spécialisé – corpus

[101] PC Pentium 4 : mémoire vive de 512Mo.

[102] La statistique sous-jacente du rapport de vraisemblance ne se prête pas à la comparaison partie-tout (Cf. ci-dessus 4.1).

entier) relève la moitié[103] des spécificités relevées par la comparaison méthodologiquement correcte (i.e. corpus spécialisé – corpus général).

	p < 0,05 toutes les fréquences
Lexico3prob5fq1	1666
AV (corpus spécialisé – corpus général)	885
AV (corpus spécialisé – corpus entier)	463
WS (corpus spécialisé – corpus général)	873
WS (corpus spécialisé – corpus entier)	455

Tableau 4.7 Nombre de spécificités positives dans les 3 outils (corpus de référence)

4.3 MÉTHODE DES MOTS-CLÉS : JUSTIFICATION

En guise de conclusion, nous proposons de procéder à une justification explicite de la méthode des mots-clés et du logiciel AV Frequency List Tool, qui permettront de déterminer les spécificités de notre corpus technique spécialisé ainsi que leur degré de spécificité. Trois arguments décisifs sont invoqués à cet effet, à savoir le type de corpus de référence, les limites techniques et la granularité des coefficients de spécificité.

Les trois outils Lexico3, Abundantia Verborum et WordSmith, sont utilisés pour l'identification des spécificités. Lexico3 se prête bien à l'étude de la distribution des mots et des unités linguistiques dans un seul corpus, divisé en plusieurs sections, et par conséquent à l'identification des spécificités dans une section particulière par rapport à l'ensemble du corpus. AV et WS, par contre, comparent un premier corpus, par exemple de langue spécialisée, à un deuxième corpus, généralement de langue générale et servant de corpus de référence, dans le but de déterminer les mots spécifiques (mots-clés) du corpus spécialisé par rapport au corpus général. Dans le cadre de notre étude, il est clair que la méthode des mots-clés (implémentée dans AV et WS) est la méthode la plus appropriée, car elle compare un corpus spécialisé à un corpus de référence de langue générale. En effet, il ne s'agit pas, dans notre étude, d'une section particulière d'un grand corpus constitué de sections similaires, mais il s'agit bel et bien d'un corpus spécialisé tout à fait indépendant du corpus de référence de langue générale.

---

[103] La fréquence (totale) des mots dans le corpus entier virtuel est modifiée également. Il n'y a plus de mots qui sont absents du corpus de référence entier et, par voie de conséquence, ces listes contiennent moins de mots-clés.

Comme nous l'avons évoqué ci-dessus, Lexico3 relève du calcul des spécificités et recourt à la distribution hypergéométrique, ce qui semble poser des problèmes techniques pour le traitement de corpus volumineux. Compte tenu de la taille importante de nos deux corpus, la distribution hypergéométrique et l'approche méthodologique du calcul des spécificités semblent moins appropriées.

L'approche globale des deux approches méthodologiques consiste à comparer des fréquences relatives dans une section ou dans un corpus spécialisé aux fréquences relatives dans l'ensemble du corpus ou dans le corpus de référence. Toutefois, la mesure ou l'échelle de la déviation est différente dans les deux approches, puisqu'on travaille avec les probabilités exactes de la distribution hypergéométrique (calcul des spécificités) dans le premier cas et avec la distribution χ^2 asymptotique du rapport de vraisemblance (LLR) (méthode des mots-clés) dans le deuxième. D'une part, le calcul des spécificités génère un exposant, le log de la probabilité, qui peut être considéré comme un coefficient de spécificité et qui est l'exposant de la base 10 pour obtenir la valeur de probabilité, exposant qui est corrélé positivement avec le degré de spécificité. D'autre part, la méthode des mots-clés calcule pour chaque mot la valeur de LLR (la mesure statistique du rapport de vraisemblance) ainsi que la valeur de probabilité correspondante. Le LLR est également corrélé positivement avec le degré de spécificité. Toutefois, l'exposant du calcul des spécificités et le LLR de la méthode des mots-clés sont deux grandeurs différentes[104].

En dépit des probabilités exactes de la distribution hypergéométrique sous-jacente, les coefficients de spécificité du calcul des spécificités (Lexico3) sont des nombres entiers arrondis. Ceux-ci sont difficiles à interpréter et à implémenter en termes de degrés de spécificité, parce qu'ils ne présentent pas de granularité très fine. Comme notre recherche vise à étudier la corrélation entre le continuum de spécificité et le continuum sémantique pour un corpus spécialisé, l'analyse des spécificités devrait conduire à un continuum de spécificité bien établi, avec des possibilités de classement précis et des degrés de spécificité avec une granularité aussi fine que possible. Il semble que Lexico3 ne se prête guère à un tel classement. Par contre, si le classement en fonction du coefficient de spécificité ne requiert pas de granularité fine, Lexico3 fournit de bons résultats, fiables.

Abundantia Verborum et l'outil AV Frequency List Tool ainsi que WordSmith Tools et son outil de mots-clés KeyWords, relèvent tous les deux de la même approche méthodologique et statistique et génèrent dès lors une liste de mots-clés ou

[104] Pour un échantillon du corpus spécialisé de 14.000 occurrences, le coefficient de spécificité (l'exposant) de Lexico3 varie entre 1 et 50 (nombres entiers), tandis que le LLR dans AV et WS varie entre 3,8 (statistiquement pertinent) et plus de 660 (nombres décimaux).

de spécificités similaire. La mesure statistique du rapport de vraisemblance (ou *keyness* dans WS), qui fait office de degré de spécificité, et la valeur p correspondante présentent une granularité beaucoup plus fine et permettent un classement plus précis. Pour des raisons pratiques de flexibilité et d'efficacité et en vue des exploitations ultérieures, nous proposons de recourir au script Python pour générer les listes de fréquences et à AV Frequency List Tool pour identifier et analyser les mots-clés ou spécificités. En effet, AV Frequency List Tool prend en entrée n'importe quelle liste de fréquence et permet donc de procéder à l'analyse des spécificités tant pour les lemmes que pour les formes graphiques (formes fléchies). WS KeyWords en revanche prend en entrée uniquement les listes de fréquence, dressées dans l'outil WS WordList, et requiert donc l'importation des deux corpus.

Dans l'annexe 7, nous procédons aux opérations permettant de dresser une liste de spécificités du corpus technique spécialisé, en le comparant au corpus de référence de langue générale. A cet effet, les deux listes de fréquence des lemmes sont requises. Comme les lemmes du corpus technique sont comparés formellement à tous les lemmes du corpus de référence, l'analyse des spécificités porte uniquement sur les unités lexicales simples. L'identification des spécificités au niveau des unités polylexicales constitue une piste de recherche à explorer ultérieurement, étant donné qu'elle requiert la mise au point d'une technique ad hoc permettant de comparer formellement toutes les unités polylexicales des deux corpus et ceci dans le but de disposer de toutes les données de la table de contingence par unité polylexicale spécifique. Cette technique de comparaison des unités polylexicales pourrait consister à identifier toutes les unités polylexicales des deux corpus (c'est-à-dire les collocations stables, pertinentes et terminologiques) et à les encoder pour pouvoir les comparer de façon univoque et automatique, ce qui constitue une contrainte opérationnelle importante de l'analyse des spécificités. Toutefois, l'analyse des unités polylexicales et la mise au point de la technique de comparaison dépassent les limites de cette thèse et feront l'objet de nos recherches ultérieures.

En conclusion, pendant cette étape de l'analyse des spécificités, les lemmes se voient attribuer un degré de spécificité et un rang de spécificité. En effet, la mesure statistique du LLR indiquant le degré de spécificité, elle permettra de classer les spécificités par ordre décroissant de spécificité et, par conséquent, de les situer sur un continuum en fonction de leur rang de spécificité. Les mots avec un degré de spécificité identique, c'est-à-dire une valeur de LLR identique, auront le même rang de spécificité. Les mots les plus spécifiques, à savoir *machine, outil, usinage, pièce, mm, vitesse, coupe,* etc. reflètent clairement la thématique du domaine du corpus technique. Pendant l'étape suivante de l'analyse des cooccurrences (Cf. chapitre 5), les spécificités du corpus technique seront soumises à la mesure de monosémie ou mesure de recoupement, qui permettra de calculer leur degré de monosémie ou degré de recoupement.

Chapitre 5

Analyse des cooccurrences

Dans le cinquième chapitre, nous expliquerons les principes méthodologiques de l'analyse des cooccurrences, qui est le deuxième axe méthodologique de notre étude. L'analyse des cooccurrences vise principalement à quantifier la monosémie et à déterminer un degré de monosémie. A cet effet, la monosémie sera implémentée comme homogénéité sémantique et elle sera étudiée à partir du recoupement des cooccurrences des cooccurrences (Cf. chapitre 2).

L'analyse du recoupement des cooccurrences des cooccurrences permettra d'attribuer un degré de monosémie aux unités lexicales spécifiques, identifiées précédemment avec la méthode des mots-clés et munies d'un degré de spécificité (Cf. chapitre 4). Dans une étape ultérieure (Cf. chapitre 7), le degré de spécificité et le degré de monosémie permettront de répondre à la question principale de la présente recherche et d'étudier la corrélation entre, d'une part, le continuum de spécificité et, d'autre part, le continuum de monosémie.

Le défi à relever dans ce chapitre méthodologique réside dans le développement d'une mesure de monosémie, qui devra permettre de quantifier la monosémie et de situer les unités lexicales analysées sur un continuum de monosémie, à l'instar du continuum de spécificité. Le développement d'une mesure de monosémie contribue également à l'étude sémantique automatisée et simultanée de plusieurs milliers d'unités lexicales, dans la mesure où on fait l'économie d'une analyse manuelle de plusieurs centaines de milliers de concordances et des contextes d'apparition pour chacune des unités lexicales analysées. Toutefois, signalons d'emblée que notre mesure de monosémie impose une restriction méthodologique importante, parce qu'elle requiert la reformulation de la monosémie traditionnelle. Cette reformulation opérationnelle consiste en l'implémentation de la monosémie en termes d'homogénéité sémantique. Par conséquent, la monosémie, telle qu'elle est étudiée dans notre étude, c'est-à-dire implémentée en termes d'homogénéité sémantique, ne correspond peut-être pas parfaitement à ce que les monosémistes traditionnels entendent par monosémie (Cf. 5.2.2.3).

Ce cinquième chapitre constitue le chapitre-clé de notre étude. Nous y préciserons d'abord les notions fondamentales de l'analyse des cooccurrences et de la désambiguïsation sémantique en général (5.1). Dans la deuxième partie (5.2), nous procéderons à un bref survol des études ayant eu recours aux cooccurrences des cooccurrences et nous expliciterons l'intérêt de ce genre d'analyse. Finalement, la troisième partie sera consacrée à la mesure de monosémie ou la mesure de recoupement (5.3), qui fera l'objet de quelques mises au point méthodologiques dans le chapitre suivant (Cf. chapitre 6).

5.1 LES COOCCURRENCES

Avant de procéder à l'analyse proprement dite, nous nous proposons d'élaborer une mesure de monosémie. Celle-ci équivaut à une mesure de recoupement, basée sur le recoupement formel des cooccurrences des cooccurrences. Le développement d'une telle mesure s'appuie sur l'analyse des cooccurrences. Dans cette perspective, on étudie le contexte linguistique ou les cooccurrences d'un mot dans le but d'identifier le sens du mot dans un contexte donné ou de déterminer les différents sens du mot.

Afin de préciser nos choix méthodologiques, nous expliquerons dans la première section les notions fondamentales de la désambiguïsation sémantique et de l'acquisition sémantique (5.1.1). Nous relèverons ensuite les aspects méthodologiques pertinents pour l'analyse des cooccurrences et, partant, pour la mesure de monosémie, tels que la fenêtre d'observation et le degré d'association (5.1.2). Nous terminerons cette partie sur les cooccurrences en abordant la notion de mesures d'association (5.1.3). Signalons d'ores et déjà que la désambiguïsation sémantique vise surtout à identifier les sens d'un mot, tandis que notre mesure cherche plutôt à attribuer au mot un degré de monosémie ou un degré d'homogénéité sémantique, ou, autrement dit, à le situer sur un continuum d'homogénéité sémantique.

5.1.1 La désambiguïsation sémantique et l'acquisition sémantique

Les travaux de désambiguïsation sémantique reposent principalement sur l'idée que le contexte, essentiellement linguistique, permet d'identifier le sens dans lequel une occurrence déterminée d'un mot ambigu est employée. Rappelons à ce propos l'adage de Firth (1957) : « You shall know a word by the company it keeps ». Le contexte permet en effet de lever ou de réduire l'ambiguïté, parce qu'il réduit l'espace des sens possibles, forçant le locuteur à tenir compte (du sens) des voisins ou des cooccurrences. Cette approche contextuelle consiste donc à « retenir pour un mot donné le sens qui se rapproche le plus de ceux de ses voisins » (Habert et al. 1997 : 108).

Alors qu'un être humain est parfaitement capable d'identifier ou de sélectionner le sens approprié d'un mot ambigu en s'appuyant sur le contexte, il n'est pas facile du tout de formaliser et d'automatiser ce processus de désambiguïsation sémantique. En matière de sémantique « machinale » (Habert et al. 2004 : 566) ou de sémantique en TAL (Traitement Automatique de la Langue), on fait généralement la distinction entre la désambiguïsation sémantique (5.1.1.1) et l'acquisition sémantique (5.1.1.2). La plupart des travaux récents se concentrent sur la désambiguïsation sémantique (Cf. Ide & Véronis 1998 ; Manning & Schütze 2002), ce qui s'explique par les nombreuses expérimentations de désambiguïsation menées dans le cadre du projet Senseval (5.1.1.3), qui vise à évaluer les systèmes de désambiguïsation.

5.1.1.1 La désambiguïsation sémantique

La désambiguïsation sémantique ou la WSD se fixe pour objectif d'associer un sens à un mot en contexte, donc d'assigner des étiquettes sémantiques (Schütze 1998). Elle procède en deux étapes. Dans un premier temps, la désambiguïsation sémantique répartit les occurrences de mots ambigus en plusieurs groupes en fonction d'un répertoire de sens préexistants (*sense discrimination*), donc sans en déterminer elle-même le sens exact. La deuxième étape, qui est l'étape principale, consiste à attribuer ou à assigner un sens (préétabli) à chaque occurrence du mot en contexte ou à chaque classe d'occurrences (*sense labeling*). Pour attribuer le sens approprié à chaque occurrence étudiée, les travaux en WSD s'appuient donc sur des sens prédéfinis, qu'ils assignent en recourant à des techniques de désambiguïsation supervisées (*corpus-based WSD*) ou à des ressources lexicales (*knowlegde-driven WSD*).

Les techniques de désambiguïsation supervisées prennent comme point de départ un corpus d'apprentissage annoté sémantiquement, afin de désambiguïser des mots ambigus dans un corpus servant de test, ce qui revient à assigner le sens approprié (*sense labelling*). Une première approche de désambiguïsation supervisée, la classification bayesienne, consiste à regarder tous les mots (pleins) autour du mot ambigu, dans une large fenêtre d'observation (*span*). Chaque informant du contexte se voit assigner la probabilité d'induire un sens et selon la règle de décision de Bayes, le modèle choisit le sens le plus probable (Manning & Schütze 2002 : 236). Toutefois, la structure et l'ordre linéaire du contexte ne sont pas pris en considération, d'où la qualification de « sac de mots » qu'on a pu attribuer à ce modèle (Manning & Schütze 2002 : 237). En plus, la présence d'un mot dans le sac est totalement indépendante de la présence d'un autre, ce qui n'est pas réaliste. L'approche des listes de décision de Yarowsky (1994) en revanche s'appuie sur des listes ordonnées d'indicateurs sémantiques (*sense informants*), obtenus sur un corpus d'apprentissage et dont les plus saillants se trouvent en tête de liste.

Bien évidemment, les techniques supervisées basées sur des corpus d'apprentissage dépendent de la disponibilité et de la fiabilité des corpus annotés sémantiquement, ce qui pose parfois problème, car l'annotation sémantique manuelle est une activité longue et fâcheuse. C'est ce qui explique les nombreux efforts d'automatisation en matière d'annotation sémantique, tels que l'emploi de corpus bilingues ou les modèles d'espace vectoriel (Schütze 1998). Le deuxième problème auquel se voient confrontées les méthodes basées sur corpus est celui de la rareté des données (*data sparseness*). En effet, il faudrait un corpus d'apprentissage très vaste pour être sûr que tous les sens d'un mot polysémique y soient représentés, à cause de la grande disparité de fréquence entre les différents sens. D'ailleurs, les multiples cooccurrences possibles d'un mot polysémique ne se retrouveront même pas dans un très vaste corpus, où elles risquent d'ailleurs d'être trop peu fréquentes pour être significatives. Dès lors, la rareté des données pose problème pour des estimations de fréquence des méthodes statistiques, basées sur des fréquences relatives de combinaisons de mots dans un corpus d'apprentissage (Ide & Véronis 1998).

Par ailleurs, les techniques de désambiguïsation ou d'apprentissage basées sur un dictionnaire ou un thésaurus utilisent des ressources lexicales supplémentaires « extérieures », tout en tenant compte des propriétés distributionnelles des sens. Les dictionnaires électroniques informatisés (*Machine-Readable Dictionaries* ou *MRD*) ont permis à Lesk (1994) de déterminer le sens d'un mot ambigu en s'appuyant sur la définition du dictionnaire ayant le plus de mots en commun avec le contexte du mot ambigu (Lesk 1994, dans Manning & Schütze 2002). Gaume et al. (2004) ont également utilisé un dictionnaire pour la désambiguïsation. Leur méthode s'appuie sur un algorithme « qui calcule une distance *sémantique* entre les mots du dictionnaire en prenant en compte la topologie complète du dictionnaire, vu comme un graphe[105] sur ses entrées » (Gaume et al. 2004 : 205). Cependant, les dictionnaires traditionnels se caractérisent souvent par un manque de cohérence et par l'absence d'informations distributionnelles, comme des contextes d'usage, des collocations et des informations syntaxiques[106]. Si les lexiques computationnels tels

[105] Un dictionnaire est considéré comme « un graphe non orienté dont les mots sont les sommets et tel qu'il existe un arc entre deux sommets si l'un apparaît dans la définition de l'autre » (Gaume et al. 2004 : 206). Un algorithme construit ensuite une mesure de similarité entre les sommets du graphe, « en rapprochant les sommets d'une même zone dense en arêtes » (Gaume et al. 2004 : 208).

[106] Des expérimentations d'étiquetage lexical et des jugements de polysémie, basés sur des dictionnaires traditionnels, aboutissent à un accord « inter-annotateur » plutôt faible de 49% (Véronis 2001 et 2004a). « Les entrées ne contiennent pas suffisamment d'indices de surface pour permettre aux annotateurs de mettre en correspondance tous les contextes avec un sens particulier de façon fiable » (Véronis 2004a : 28).

que WordNet fournissent des définitions, des sets de synonymes et des relations sémantiques, la granularité des sens de WordNet est souvent trop fine pour la désambiguïsation sémantique[107] (Ide & Véronis 1998).

Finalement, les travaux basés sur des thésaurus s'appuient sur l'idée que les catégories du thésaurus, qui sont utilisées comme des approximations de classes conceptuelles, correspondent à des distinctions de sens. Ainsi, le mot *grue* (anglais : *crane*) relève de la catégorie des animaux (« grand oiseau ») et de la catégorie des machines (« machine de levage »). Comme les classes conceptuelles différentes apparaissent dans des contextes clairement différents, les indicateurs contextuels d'une catégorie (par exemple *moteur, piston, engrenage* pour la catégorie des machines) pourront aussi servir d'indicateurs sémantiques pour les membres de cette catégorie, en l'occurrence *grue* au sens de « machine de levage » (Yarowsky 1992). Citons à ce sujet l'approche de Yarowsky basée sur la catégorisation sémantique du *Roget's International Thesaurus* (Yarowsky 1992).

5.1.1.2 L'acquisition sémantique

L'acquisition sémantique comprend trois volets : (1) « la mise en évidence de relations sémantiques entre mots par leur cooccurrence dans des contextes lexicosyntaxiques spécifiques » (Habert et al. 2005 : 278), (2) « la mise en évidence de similarités sémantiques entre mots à partir de distributions proches » (ibid.) et (3) « la caractérisation des différentes acceptions d'un mot » (ibid.), dénommée aussi *sense induction* (Yarowsky 1995) ou *word sense discrimination* (Schütze 1998), qui vise principalement la discrimination ou le « dégroupement » de sens (Dorow & Widdows 2003). L'acquisition sémantique consiste donc à rechercher des similarités (ou des dissimilarités) sémantiques à partir de similarités (ou de dissimilarités) distributionnelles, à l'aide du contexte linguistique. Le but de l'acquisition sémantique est double. D'une part, le dégroupement (ou la division) des sens consiste à repérer les cas où un mot est employé « simultanément avec des sens différents au sein d'un corpus » (Habert et al. 2005 : 278). Les contextes d'emploi de ces « mots aux sens mouvants » sont souvent très différenciés, par exemple *grêve* « arrêt du travail » et « plage de gravier » ou *guerre* dans *guerre aérienne* et dans *guerre médiatique*. D'autre part, le regroupement (ou *clustering*) des occurrences d'un mot ambigu en groupes ou clusters vise à déterminer quelles occurrences du mot ont le même sens. Les similarités sémantiques sont mises en évidence à partir de distributions proches. Il est à noter que ces groupes d'occurrences ne

---

[107] Il est à noter que la réduction des « étiquettes fournies par les annotateurs aux seules divisions de plus haut niveau dans la hiérarchie des entrées » (Véronis 2004a : 28) ne permet pas d'améliorer significativement l'accord inter-annotateur (Véronis 2004a).

correspondent pas nécessairement à des subdivisions sémantiques standard (Yarowsky 1995). En effet, l'acquisition sémantique n'a pas recours à des sens préexistants ou préétablis, contrairement à la désambiguïsation sémantique.

Habert et al. (2004 et 2005) soulèvent une série de problèmes techniques et théoriques en matière d'acquisition sémantique, qui expliquent notamment pourquoi les travaux récents explorent surtout les regroupements d'occurrences et les similarités sémantiques. D'abord, du point de vue technique, pour les outils de traitement, « il s'agit toujours de la même chaîne de caractères », bien que dégrouper les sens consiste à « trouver le moyen de repérer les cas où un mot en cache un, voire plusieurs autre(s) » (Habert et al. 2004 : 566). Ensuite, du point de vue théorique, l'acquisition sémantique souffre de la prépondérance de la vision sémantique fixiste et discrétisante, selon laquelle les sens d'un mot sont discrets et totalement disjoints (Habert et al. 2004).

A défaut de corpus d'apprentissage annotés sémantiquement ou à défaut de ressources lexicales appropriées, le recours à des techniques de désambiguïsation ou d'apprentissage non supervisées s'impose. Ainsi, dans les domaines spécialisés, les dictionnaires et thésaurus généraux s'avèrent peu utiles et les sources d'information externes ou les jugements humains ne sont pas toujours disponibles. Comme on ne connaît pas au préalable la classification des données, la désambiguïsation non supervisée revient à une tâche de regroupement ou d'agglomération (*clustering*). Citons notamment les travaux de Schütze (1998) sur la *context-group discrimination*, où la désambiguïsation se fait par un calcul de probabilité de chaque sens, à partir des mots figurant dans le contexte et par la décomposition en valeurs singulières (*Singular Value Decomposition* ou SVD[108]). Schütze ne recourt pas à des listes de sens préétablis, mais extrait automatiquement du corpus la liste des sens ou des « usages ». « Ces usages correspondent à des groupes (*clusters*) de contextes similaires dans un espace de très grande dimensionnalité formé par des vecteurs de mots ou de cooccurrences proches du mot à désambiguïser » (Véronis 2003 : 266). La décomposition en valeurs singulières permet de réduire la dimensionnalité de l'espace et de faire émerger les différents groupes de contextes similaires, définis de

---

[108] La décomposition en valeurs singulières est le principe méthodologique fondamental de l'analyse sémantique latente (*Latent Semantic Analysis* ou LSA) (Landauer, Foltz & Laham 1998 ; Landauer 2002). L'analyse sémantique latente est un modèle de représentation vectoriel de la signification des mots. Chaque mot est représenté par un vecteur dans un espace de plusieurs centaines de dimensions et le degré d'association entre deux mots est déterminé par le cosinus de leur angle. Le but de l'analyse sémantique latente est de produire des valeurs d'association entre les mots, par la réduction des dimensions de la matrice des occurrences de ces mots (Denhière & Lemaire 2003) (Cf. 5.2.).

manière distributionnelle, et dès lors les différents usages ou sens. Ce type de désambiguïsation relève des méthodes de regroupement dur (*hard clustering*) ou de regroupement en classes disjointes. Par contre, l'approche adoptée par De Marneffe et Dupont (2004) relève du regroupement en classes non disjointes (*soft clustering*) et vise à inclure des informations linguistiques pour améliorer les approches statistiques. Toutefois, elle prend comme point de départ une distribution normale (gaussienne), difficilement compatible avec les données linguistiques étudiées.

Il est clair que notre étude s'inscrit plutôt dans la perspective méthodologique de l'acquisition sémantique, étant donné qu'elle consiste à repérer des similarités sémantiques à partir de similarités distributionnelles. En effet, notre analyse vise à étudier les cooccurrences de toutes les occurrences d'un mot potentiellement ambigu et à vérifier dans quelle mesure ces cooccurrences sont sémantiquement apparentées. Finalement, le but de l'étude est d'évaluer si les occurrences du mot ambigu sont sémantiquement homogènes ou hétérogènes. Notons que nous ne cherchons aucunement à établir des groupes d'occurrences nettement délimités, d'autant moins que les groupes d'occurrences semblent se caractériser en général par des frontières floues (Habert et al. 2004). Compte tenu de ces observations et de nos objectifs de recherche, l'idée d'un continuum sémantique ou, plus précisément, d'un continuum d'homogénéité sémantique nous paraît plus appropriée (Cf. 5.3).

5.1.1.3 Evaluation des systèmes de désambiguïsation

Senseval[109] est une organisation internationale qui s'est fixée pour but d'évaluer les systèmes et programmes de désambiguïsation lexicale et sémantique automatique, principalement en anglais (Kilgarriff & Palmer 2000). Pour le français et l'italien, le sous-groupe Romanseval[110] adopte les mêmes principes de base (Véronis 1998 ; Segond 2000). Les participants au projet travaillent tous sur le même corpus (à peu près un million d'occurrences par langue), constitué de questions écrites posées par des parlementaires européens sur des sujets variés tels que l'environnement, l'économie, l'éducation, etc. (Véronis 1998). Les participants sont censés désambiguïser et étiqueter une liste de 60 mots ambigus (substantifs, adjectifs, verbes), tels que *barrage, constitution, simple, arrêter* (Segond 2000). Comme ils disposent également de la liste des différents sens potentiels des 60 mots ambigus, la tâche consiste principalement à assigner le sens approprié aux occurrences des mots à désambiguïser, en se basant sur des indices contextuels.

---

[109] Http://www.senseval.org.

[110] Http://www.lpl.univ-aix.fr/projects/romanseval/.

Un corpus d'évaluation, annoté sémantiquement par des annotateurs humains, permet d'évaluer les résultats des différents programmes de désambiguïsation en termes de précision (*precision*) et de rappel (*recall*). La précision indique le nombre de réponses pertinentes ou correctes par rapport au nombre total de réponses données. Le rappel est le rapport entre le nombre de réponses correctes données et le nombre total de réponses correctes possibles. Les résultats des participants sont donc comparés à un « *Gold Standard* », indiquant pour chaque occurrence des mots à désambiguïser le sens attribué par les annotateurs humains, en utilisant le principe de l'accord inter-annotateur (*interannotator agreement*). Les initiatives Senseval et Romanseval ont donné lieu à la publication de nombreux articles sur les systèmes et programmes de désambiguïsation et sur les résultats obtenus (Ellman, Klincke & Tait 2000 ; Lin 2000 ; Segond, Aimelet et al. 2000 ; Suderman 2000 ; Yarowsky 2000), ainsi que sur des questions méthodologiques fondamentales en matière de désambiguïsation (Cf. Hanks 2000 ; Ide 2000).

Les systèmes et algorithmes de désambiguïsation et les mesures développées dans le cadre de Senseval et Romanseval visent donc à attribuer le sens approprié à des occurrences de mots ambigus et à améliorer les performances de désambiguïsation en termes de précision et de rappel. Par conséquent, il est difficile de comparer notre mesure de monosémie à ces mesures et algorithmes de désambiguïsation. La mesure que nous développons dans le cadre de notre étude ne cherche pas à attribuer des sens préétablis, mais à déterminer les caractéristiques sémantiques des mots ambigus ou non, dans le but de déterminer leur degré d'homogénéité sémantique. Les algorithmes d'acquisition sémantique, qui visent le regroupement des occurrences de mots ambigus en clusters, tels que la technique de Schütze (1998) ou la LSA, pourraient éventuellement se prêter à une comparaison avec notre mesure de monosémie. Toutefois, ces algorithmes sont des modèles plutôt aveugles, étant donné qu'ils ne tiennent pas compte des caractéristiques (sémantiques, syntaxiques, etc.) des cooccurrences. Pour le développement de notre mesure, nous nous proposons d'inclure certaines caractéristiques des cooccurrences, afin d'enrichir la mesure de monosémie et de la rendre plus précise.

5.1.2 Aspects méthodologiques pertinents

Il est communément admis que les cooccurrences « constituent de forts indices désambiguïsateurs pour distinguer les différents usages des mots » (Véronis 2003 : 266). Toutefois, il convient de se pencher sur un certain nombre de questions méthodologiques. La question se pose notamment de savoir s'il faut prendre en considération toutes les cooccurrences qui apparaissent avec le mot de base (mot-cible à désambiguïser ou à caractériser sémantiquement) ou uniquement les cooccurrences privilégiées. Et comment déterminer leur degré d'association et jusqu'à quelle distance du mot-cible ou du mot à désambiguïser faut-il aller ?

Etant donné que les recherches en désambiguïsation et acquisition sémantique s'appuient principalement sur l'analyse de cooccurrences et qu'elles privilégient l'axe syntagmatique, elles nous permettent de relever un certain nombre d'aspects méthodologiques pertinents pour le développement de notre mesure de monosémie. Les cooccurrences se prêtent non seulement à l'identification du sens d'un mot polysémique en contexte ou à la sélection du sens approprié dans une liste de sens préétablis. Elles s'avèrent également indispensables pour regrouper les occurrences synonymiques et pour déterminer si les occurrences sont sémantiquement homogènes ou hétérogènes, et à quel point.

5.1.2.1 L'approche « sac de mots »

Ide et Véronis (1998) font la distinction entre deux approches du contexte, selon la prise en compte ou non des relations entre le mot à désambiguïser et son contexte (ou ses cooccurrences) : l'approche « sac de mots » et l'approche de « l'information relationnelle ». La première approche consiste à considérer tous les mots (ou tous les mots pleins) dans une certaine fenêtre d'observation autour du mot à désambiguïser, mais sans tenir compte de l'ordre linéaire des mots entre eux ni des relations. Mentionnons en guise d'exemple les travaux de Schütze (1998) ou l'analyse sémantique latente (LSA) (Landauer, Foltz & Laham 1998 ; Landauer 2002), qui recourent à la décomposition en valeurs singulières (SVD) à partir d'ensembles non ordonnés de mots. Or, en réalité les mots n'apparaissent pas indépendamment les uns des autres : certaines combinaisons de mots sont bien plus probables que d'autres, certaines associations de mots sont plus fortes que d'autres et il convient de tenir compte de ces informations (Cf. 5.1.2.2).

La deuxième approche, celle de l'information relationnelle, insiste sur l'importance des relations entre le mot à désambiguïser et son contexte. Dans la fenêtre d'observation (*span*), on tient compte des relations syntaxiques, des préférences de sélection et des collocations (Ide & Véronis 1998). Selon Audibert (2003), il y aurait une baisse significative des performances de désambiguïsation, si l'on ne tient pas compte de la position ou de la distance des cooccurrences par rapport au mot à désambiguïser. Selon Yarowsky (2000), la prise en compte de la classe lexicale des cooccurrences et de leur relation syntaxique par rapport au mot à désambiguïser permet également d'aboutir à des précisions importantes (Yarowsky 2000). En plus, la combinaison de plusieurs sources linguistiques permet d'améliorer les résultats de la désambiguïsation. Citons parmi ces sources les étiquettes syntaxiques (*POS-tags* ou *part-of-speech tags*) (Cf. 5.1.2.3), des informations de fréquence, des informations morphologiques, des collocations et des associations entre mots et contexte sémantique (*clusters*) (Stevenson & Wilks 2001).

Les associations de mots ou les cooccurrences significatives désignent le phénomène par lequel deux mots sont utilisés dans le même contexte linguistique (c'est-à-dire dans la même fenêtre d'observation) plus souvent que par hasard (associations arbitraires) ou plus souvent qu'on ne s'y attendrait en fonction de leurs fréquences globales dans les autres contextes du corpus. Les « collocations[111] » sont des cooccurrences restreintes à des associations de mots liés grammaticalement (Manning & Schütze 2002). Les collocations, par exemple *célibataire endurci*, se caractérisent donc par la rigidité syntaxique (plus ou moins grande), mais aussi par l'irrégularité sémantique et par le fait qu'elles constituent une unité syntaxique et sémantique. La récurrence (ou la co-fréquence élevée) résulte principalement du processus de lexicalisation. Les principales caractéristiques des collocations correspondent aux principes de non-compositionnalité (impossible ou difficile de calculer le sens de la collocation à partir des composantes), de non-substituabilité (impossible ou difficile de substituer des synonymes aux composantes de la collocation) et de non-modifiabilité (impossible ou difficile de modifier la collocation par des éléments lexicaux supplémentaires ou par des transformations grammaticales) (Manning & Schütze 2002).

De ce qui précède, il ressort que l'approche « sac de mots » n'est pas l'approche idéale pour la désambiguïsation ou l'analyse sémantique et qu'il est important, au contraire, de tenir compte également des relations entre le mot à désambiguïser et ses cooccurrences. Evidemment, la prise en compte de ces relations et des informations d'association des cooccurrences est tributaire de l'annotation du corpus et se fera en fonction des objectifs de recherche.

5.1.2.2 Le degré d'association

Plus les degrés d'association sont élevés, plus les combinaisons de mots qui en résultent sont idiomatiques. Si les collocations et les cooccurrences statistiquement significatives, avec un degré d'association élevé, apparaissent ensemble fréquemment, cela ne peut être dû au hasard. Ces cooccurrences significatives sont donc très importantes pour la désambiguïsation ou pour l'analyse sémantique d'un mot de base, car elles contiennent des indications sémantiques précieuses. Mentionnons à ce sujet l'hypothèse d'« un sens par collocation » (*one sense per collocation*) (Yarowsky 1995). Dans une collocation, un mot serait utilisé dans un seul sens avec 90-99% de probabilité. Les mots voisins et les cooccurrences pertinentes sont des indices importants quant au sens du mot ambigu, si l'on tient

[111] Selon la terminologie de Haussmann (1979), une collocation se compose d'une base (mot de base, mot-cible ou *node* en anglais) et d'un collocatif (cooccurrent ou *collocate* en anglais). C'est l'approche classique où la collocation est composée de deux mots.

compte entre autres de la distance relative, de l'ordre des mots et de leur relation syntaxique (Yarowsky 1995). La prise en considération des informations sur le degré d'association entre un mot et ses cooccurrents s'avère donc tout à fait utile. A cet effet, plusieurs mesures d'association permettent de déterminer les cooccurrences significatives ainsi que leur degré d'association (Cf. 5.1.3).

5.1.2.3 Catégorie grammaticale et mots grammaticaux

Il est peut-être intéressant d'inclure aussi des informations concernant la catégorie grammaticale des cooccurrences. A ce sujet, Yarowsky (1992) signale que les verbes sont le mieux désambiguïsés par leur COD, les substantifs par les adjectifs et les substantifs adjacents (Cf. 5.1.2.4) et les adjectifs par les substantifs qu'ils modifient. Audibert (2003) fait état d'observations similaires et insiste sur le rôle des mots grammaticaux. Il serait en effet plus judicieux d'inclure des prépositions pour désambiguïser les substantifs et des pronoms personnels pour les verbes. En général, le retrait des mots grammaticaux de l'algorithme de désambiguïsation risque d'entraîner une baisse des performances (Audibert 2003). Selon Suderman (2000), les mots les plus pertinents sont les mots immédiatement adjacents, y compris les mots grammaticaux, dans une fenêtre d'observation de taille limitée.

5.1.2.4 La fenêtre d'observation

Les mots adjacents, les voisins qui précèdent et qui suivent, tant des mots lexicaux que grammaticaux, s'avèrent les plus pertinents pour la désambiguïsation et pour l'analyse sémantique en général. Dès lors, il convient de se poser des questions sur la distance idéale ou sur la taille de la fenêtre d'observation. « Deux unités sont cooccurrentes si elles figurent ensemble dans une unité de contexte (le voisinage) » (Habert et al. 1997 : 192). Cette unité de contexte pourrait notamment se définir par les *k* mots avant et par les *k* ou *l* mots après le mot de base, qui détermineraient alors la taille de la fenêtre d'observation.

La prise en compte du contexte, le plus souvent dans une fenêtre d'observation déterminée, comprend les informations du micro-contexte, du contexte topique et du domaine. Le micro-contexte ou le contexte local se situe dans une petite fenêtre de mots avoisinants (variant de quelques mots à toute la phrase). Selon Yarowsky (1994), la distance idéale pour les ambiguïtés syntaxiques est une fenêtre d'observation de 3 ou 4 mots. Par contre, pour les ambiguïtés sémantiques, qui dépendent du sujet (*topic-based*), il propose une fenêtre beaucoup plus large de 20 à 50 mots (Yarowsky 1994). Le contexte topique plus large comprend généralement quelques phrases et exploite la redondance dans le texte (Ide & Véronis 1998). Gale, Church & Yarowsky (1993) observent environ 50 mots autour du mot polysémique. Signalons à cet effet l'hypothèse d'« un sens par discours » (Yarowsky 1995) : dans

un discours déterminé, des mots ambigus seraient utilisés dans un seul sens avec un degré assez élevé de probabilité.

Le domaine permettrait également de désambiguïser, dans la mesure où seul le sens pertinent par rapport aux domaine serait activé. Les limitations de cette approche sont évidentes[112] : le domaine n'élimine pas l'ambiguïté de tous les mots. De ce point de vue, l'hypothèse de Yarowsky mentionnée ci-dessus est discutable parce que l'influence du domaine dépend de plusieurs facteurs, notamment du type de texte (degré de technicité du texte) et de la relation entre les sens du mot ambigu (fortement polarisés, usage spécialisé, etc.) (Ide & Véronis 1998).

La fenêtre d'observation ou la taille du contexte à prendre en considération dépend principalement des objectifs de recherche et des caractéristiques formelles et opérationnelles du corpus d'analyse. Un contexte large, tel que préconisé par Yarowsky (1995), augmente les risques de bruit et d'indications sémantiques non pertinentes. Une taille similaire de la fenêtre d'observation se retrouve dans l'approche de Schütze (1998), à savoir 25 mots à gauche et 25 mots à droite, mais la technique de décomposition en valeurs singulières permet de filtrer et d'éliminer les informations non pertinentes (le bruit). Audibert (2003) quant à lui obtient les meilleurs résultats de désambiguïsation dans une fenêtre de 1 à 4 mots autour du mot à désambiguïser. Bien évidemment, des fenêtres trop étroites ne permettent pas de retrouver toutes les cooccurrences ni les informations sémantiques pertinentes et risquent donc de se heurter au problème du silence. Les fenêtres d'observation les plus courantes s'étendent de 3 à 5 mots autour du mot à désambiguïser (Suderman 2000 ; de Loupy, El-Bèze & Marteau 2000 ; Weber, Vos & Baayen 2000 ; Lapata 2002 ; Habert et al. 2005).

5.1.2.5 La lemmatisation

On peut se demander s'il faut considérer les cooccurrences au niveau des formes fléchies ou plutôt au niveau des formes canoniques. Les études d'acquisition sémantique, qui recourent à des matrices et à des techniques automatiques de décomposition en valeurs singulières, analysent les cooccurrences au niveau des formes fléchies (Schütze 1998 ; Karov & Edelman 1998). Les autres études de désambiguïsation sémantique (Yarowsky 1992 et 1994 ; Stevenson & Wilks 2001 ; Lapata 2002 ; Audibert 2003) utilisent plutôt les lemmes des cooccurrences. Lapata (2002) préfère même les lemmes à l'utilisation d'étiquettes syntaxiques (*POS-tags*).

---

[112] « *The lawyer stopped at the bar for a drink* » (Ide & Véronis 1998 : 22). Dans un document juridique, le mauvais sens serait activé (« barreau » au lieu de « débit de boissons »).

5.1.2.6 La pondération

Il peut être intéressant aussi de procéder à une pondération des cooccurrences. Cette pondération pourrait s'envisager en fonction de plusieurs facteurs, notamment la fréquence, la distance du mot à désambiguïser et l'étiquette syntaxique (ou la catégorie grammaticale) (Karov & Edelman 1998). Elle se justifie pour plusieurs raisons. Premièrement, les cooccurrences plus fréquentes apportent généralement moins d'informations sur le sens et sur la similarité de sens (Karov & Edelman 1998). Dès lors, elles seront moins importantes pendant la désambiguïsation et se verront attribuer un poids moins lourd. Cependant, cette affirmation semble contredire l'importance de la prise en considération des mots grammaticaux, qui sont très fréquents[113]. Deuxièmement, les mots qui se trouvent plus loin du mot ambigu apportent moins d'informations, d'où l'importance d'une fenêtre d'observation limitée, telle que 5 mots autour du mot ambigu. En conclusion, Karov et Edelman (1998) tiennent compte de l'étiquette syntaxique et envisagent un poids différent pour les noms (1,0), les verbes (0,6) et les adjectifs (0,1).

Une pondération permettrait effectivement de tenir compte du pouvoir désambiguïsateur différent des cooccurrences du mot de base, non seulement en fonction de la catégorie grammaticale, mais également en fonction de la fréquence ou de la distance. Il serait également envisageable d'opérer une pondération en fonction du degré d'association ou même en fonction de la saillance des cooccurrences, telle qu'elle est implémentée dans le modèle des catégories du thésaurus (Yarowsky 1992). L'idée d'intégrer la saillance (ou la spécificité) des cooccurrences nous paraît particulièrement intéressante pour préciser la mesure de monosémie (Cf. 5.3 et chapitre 6).

5.1.3 Les mesures d'association

Les mesures statistiques d'association pour identifier les collocations et les cooccurrences significatives s'appuient toutes sur l'adage de Firth[114], soit sur la proximité lexicale, c'est-à-dire sur les cooccurrences du mot de base ou sur son contexte linguistique. Comme nous l'avons évoqué ci-dessus (Cf. 5.1.2.2), le degré d'association entre le mot de base (mot-cible) et ses cooccurrents est un aspect méthodologique très important pour la désambiguïsation sémantique et pour l'analyse sémantique quantitative. En effet, le degré d'association ou la

---

[113] Les prépositions permettent de désambiguïser, par exemple *il cède* versus *il cède quelque chose à quelqu'un*.

[114] « You shall know a word by the company it keeps » (Firth 1957).

significativité de cooccurrence (*collocative significance*) permet de quantifier la relation entre le mot de base et ses cooccurrents ou voisins. Le degré d'association est calculé à partir des fréquences observées et attendues d'une paire de mots, à l'aide d'une mesure d'association. Les fréquences observées (O) et les fréquences attendues (E) (*expected frequencies*) sont généralement exprimées dans une table de contingence[115] (Cf. tableaux 5.1 et 5.2). La co-fréquence observée est la fréquence d'occurrence de la paire mot1+mot2 (*poser* + *question*), c'est-à-dire la fréquence totale de toutes les occurrences du mot1 avec les occurrences du mot2. La co-fréquence observée est exprimée par O_{11}. Le nombre total d'occurrences dans le corpus est exprimé par N ($N = O_{11}+O_{12}+O_{21}+O_{22}$) (Cf. tableau 5.1). Les fréquences par rangée (R_1, R_2) ou par colonne (C_1, C_2) sont qualifiées de fréquences marginales. Le total des fréquences des rangées et des colonnes équivaut à N.

	Mot2 = *question*	Mot2 ≠ *question*	Fréquence par rangée
Mot1 = *poser*	O_{11} (= co-fréquence) *poser une question*	O_{12} p.ex. *poser un diagnostic*	R_1 $= O_{11} + O_{12}$
Mot1 ≠ *poser*	O_{21} p.ex. *répondre à une question*	O_{22} p.ex. *répondre à une annonce*	R_2 $= O_{21} + O_{22}$
Fréquence par colonne	C_1 $= O_{11} + O_{21}$	C_2 $= O_{12} + O_{22}$	N $(= O_{11}+O_{12}+O_{21}+O_{22})$

Tableau 5.1 Table de contingence : fréquences observées

Si la co-fréquence observée O_{11} (Cf. tableau 5.1) ou la fréquence de cooccurrence des deux mots (mot de base + cooccurrent) dépasse la co-fréquence attendue E_{11} (Cf. tableau 5.2), compte tenu des fréquences individuelles des deux mots dans le corpus, l'association récurrente mot1+mot2 (mot de base + cooccurrent) est statistiquement significative.

	Mot2 = *question*	Mot2 ≠ *question*
Mot1 = *poser*	E_{11} $= (R_1C_1)/N$	E_{12} $= (R_1C_2)/N$
Mot1 ≠ *poser*	E_{21} $= (R_2C_1)/N$	E_{22} $= (R_2C_2)/N$

Tableau 5.2 Table de contingence : fréquences attendues

115 Http://www.collocations.de/AM/index.html.

Pour identifier les cooccurrences récurrentes significatives, différentes mesures statistiques d'association sont disponibles. La plupart de ces approches et mesures statistiques prennent comme point de départ la question de savoir si la cooccurrence des deux mots est arbitraire ou, par contre, si les deux mots apparaissent ensemble plus souvent que par hasard. Cette question est reformulée sous forme de l'hypothèse nulle d'indépendance des deux mots (cooccurrence ou association arbitraire). Si la probabilité (valeur p) de cooccurrence sous l'hypothèse nulle d'indépendance est très faible et inférieure à un seuil de significativité déterminé (par exemple une valeur $p < 0,05$ ou $p < 0,01$ ou même $p < 0,001$)[116], l'hypothèse nulle est rejetée et la cooccurrence est statistiquement significative (Manning & Schütze 2002). Les probabilités et les degrés d'association sont calculés à l'aide d'une mesure d'association, permettant non seulement de repérer les cooccurrences statistiquement significatives mais également d'ignorer le bruit, c'est-à-dire les associations ou combinaisons arbitraires (*random*) (Evert & Krenn 2003).

Récemment, de nombreuses études et recherches se sont penchées sur les différentes mesures d'association, ainsi que sur leurs caractéristiques et leurs diverses performances (Weber, Vos & Baayen 2000 ; Evert & Krenn 2001 ; Manning & Schütze 2002 ; Evert & Krenn 2003 ; Evert & Kermes 2003 ; Pezik 2005). Les mesures d'association couramment[117] utilisées sont l'information mutuelle (*Mutual Information* ou MI), le test t (*t-test*), le score Z (*Z-score* ou écart-réduit), le χ^2, la mesure statistique du rapport de vraisemblance (G^2 ou LLR) et le test de Fisher Exact (calcul hypergéométrique).

La plupart de ces mesures statistiques se prêtent aussi bien à l'identification des cooccurrences que des spécificités (Cf. chapitre 4). Certaines mesures statistiques d'association relèvent de tests exacts, d'autres de tests asymptotiques (approximatifs). D'autres mesures encore ne sont pas basées sur des tests statistiques d'hypothèse, mais sur des combinaisons heuristiques de fréquences observées et marginales. La co-fréquence (ou la fréquence observée de cooccurrence) est souvent utilisée comme point de référence (*baseline*) pour la comparaison et pour l'évaluation de différentes mesures d'association (Cf. Evert & Krenn 2001).

---

[116] En linguistique computationnelle et en TAL, les seuils sont souvent très sévères (0,001), en raison de la quantité importante de données textuelles (Manning & Schütze 2002).

[117] Nous nous limiterons dans notre thèse aux mesures d'association fréquemment utilisées pour la détection de cooccurrences significatives. Pour une comparaison plus approfondie, voir : http://www.collocations.de/AM/index.html.

Il est à noter que les résultats de la plupart des mesures d'association ne se prêtent pas à une comparaison directe, c'est-à-dire en termes de degrés d'association absolus, mais plutôt à une comparaison des cooccurrences repérées et de leur classement (*ranking*) à partir des degrés d'association.

5.1.3.1 Mesures basées sur les fréquences observées et marginales

Les coefficients de Dice[118] et de Jaccard[119] sont basés sur des proportions de fréquences observées et marginales (Cf. tableau 5.1).

L'information mutuelle ou MI (Church & Hanks 1990) relève de la théorie de l'Information (*Information Theory*) et compare la probabilité de co-fréquence de deux mots avec la probabilité d'apparition indépendante de chaque mot. S'il y a une vraie association privilégiée (*genuine association*) entre les deux mots, la probabilité jointe $P(x,y)$ sera beaucoup plus importante que les deux probabilités indépendantes $P(x)\cdot P(y)$[120] ou beaucoup plus importante que la chance. Cependant, la mesure de l'information mutuelle a tendance à surestimer l'association de paires de mots peu fréquentes (Weber, Vos & Baayen 2000 ; Manning & Schütze 2002) et dès lors, elle est moins appropriée pour des cooccurrences rares, surtout lorsque la co-fréquence attendue E_{11} est très limitée (Evert & Krenn 2003). En dépit de ce risque de surestimation, la mesure statistique d'association de l'information mutuelle est souvent utilisée en lexicographie.

5.1.3.2 Mesures basées sur des tests exacts

Les tests statistiques exacts calculent la probabilité totale (valeur p) qu'on a d'observer des fréquences similaires (ou supérieures) aux fréquences observées. Si cette probabilité est très faible et inférieure à un seuil de significativité déterminé, l'hypothèse nulle sera rejetée. Notons que l'hypothèse nulle des tests exacts est valable uniquement pour un échantillon (Evert 2002). Des valeurs p très faibles indiquent des cooccurrences très significatives et donc des associations très fortes.

---

[118] Dice = $\frac{2 \cdot O_{11}}{R_1 + C_1}$ (= moyenne harmonique) (Evert & Krenn 2003).

[119] Jaccard = $\frac{O_{11}}{O_{11} + O_{12} + O_{21}}$ (Evert & Krenn 2003).

[120] $I(x,y) = \log_2 \frac{P(x,y)}{P(x) \cdot P(y)}$ (Church & Hanks 1990) ou MI = $\log \frac{O_{11}}{E_{11}}$ (Evert & Krenn 2003).

Comme les probabilités calculées sont souvent extrêmement faibles, l'utilisation d'un logarithme négatif en base 10 (Cf. chapitre 4) permet d'obtenir une échelle plus commode et plus facilement interprétable, où des valeurs élevées indiquent des degrés d'association élevés.

Le test de Fisher Exact (calcul hypergéométrique)[121] calcule des probabilités exactes et dès lors, cette mesure exacte convient très bien à des cooccurrences peu fréquentes ou à des corpus peu volumineux. Dans des corpus plus volumineux, il est possible de recourir à une approximation binomiale ou à une approximation poissonienne (Cf. chapitre 4).

5.1.3.3 Mesures basées sur des tests asymptotiques

Contrairement aux tests exacts, les tests asymptotiques ne calculent pas des probabilités exactes, mais des statistiques de test, qui donnent une indication approximative d'une distribution connue, pour $N \rightarrow \infty$ (Evert 2002). Les tests asymptotiques permettent ainsi de remédier principalement au problème de la complexité numérique des tests exacts. La statistique de test de ces mesures d'association indique le degré d'association et la valeur p correspondante se prête facilement à une comparaison avec la valeur p des tests exacts.

- *Le score Z et le test t*

La mesure du score Z[122] (*Z-score*) est la version asymptotique de la mesure binomiale et permet d'atteindre de façon approximative les distributions discrètes (les distributions binomiale ou poissonienne) à l'aide d'une distribution continue (la distribution normale). Tout comme les autres mesures d'association, le score Z sert à repérer des associations de mots récurrentes et pertinentes. Néanmoins, lorsque la co-fréquence attendue (E_{11}) est limitée, sous l'hypothèse nulle, le score Z gonfle le degré d'association des associations de mots peu fréquentes. En raison de cette surestimation, le score Z n'est pas fiable pour les fréquences faibles et les cooccurrences rares[123].

---

[121] Pour les explications détaillées sur le calcul hypergéométrique : voir le chapitre précédent (Cf. 4.1.1).

[122] Z-score = $\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$ (Evert & Krenn 2003).

[123] La correction de Yates essaie de remédier au problème des erreurs d'estimation de la distribution normale et propose le numérateur suivant dans la formule de Z-score $|O_{11}-E_{11}|-0,5$ (Evert 2002).

Le résultat du test t[124], appelé le score t (*t-score*), ressemble beaucoup au score Z, dans la mesure où il suppose aussi une distribution normale des probabilités. Toutefois, le test t évite le problème des faibles fréquences du score Z, en raison de l'adaptation du dénominateur de la formule qui indique la variance ($\sqrt{O_{11}}$ au lieu de $\sqrt{E_{11}}$). Comme nous l'avons évoqué ci-dessus (Cf. chapitre 4), la distribution normale n'est pas compatible avec l'analyse de corpus linguistiques.

- *Test du chi-carré (χ²) de Pearson*

Le test du chi-carré (χ²) de Pearson permet d'évaluer l'indépendance des valeurs d'une table de contingence, telle qu'une table 2x2 (Cf. tableau 5.1). La statistique de test a une distribution χ² asymptotique (approximative) avec 1 degré de liberté (*df*)[125] pour une table de contingence de 2x2.

Le test du chi-carré (χ²) de Pearson consiste à comparer les fréquences observées aux fréquences attendues, sous l'hypothèse nulle d'indépendance. Si la différence entre les fréquences observées et les fréquences attendues est suffisamment grande, l'hypothèse nulle pourra être rejetée et la co-fréquence observée sera statistiquement significative. Comme le test du chi-carré est un test bilatéral, tant les différences positives que négatives seront prises en compte. Pour faciliter la comparaison avec les résultats d'autres mesures d'association, il vaut mieux la convertir en un test unilatéral et rejeter l'hypothèse nulle lorsque $O_{11} > E_{11}$ (Evert & Krenn 2003). Le test du chi-carré (χ²) de Pearson avec la correction de Yates reprend dans le numérateur -0,5. Toutefois, le test du chi-carré (χ²) a toujours tendance à surestimer le degré d'association des associations de mots rares, c'est-à-dire peu fréquentes, et dès lors, il convient moins bien dans des situations de rareté des données[126].

- *Rapport de vraisemblance (Log-Likelihood Ratio)*

Les rapports de vraisemblance se prêtent mieux au problème de rareté des données et donc aux associations de mots peu fréquentes. Le rapport de vraisemblance ou le rapport de probabilité est le rapport entre, d'une part, la probabilité maximale sous

---

[124] T-score = $\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$ (Evert & Krenn 2003).

[125] *Df* = (nombre de rangées -1)·(nombre de colonnes -1). Pour une table de contingence 2x2, le nombre de degrés de liberté est égal à (2-1)·(2-1) = 1·1 = 1.

[126] Lorsque les associations de mots ou les cooccurrences sont trop peu fréquentes, il y a un problème de manque de données et, par conséquent, les résultats ou les estimations des tests statistiques ne sont pas fiables.

l'hypothèse nulle d'indépendance (l'association arbitraire) et, d'autre part, la probabilité maximale de dépendance. La mesure statistique du rapport de vraisemblance, G² ou LLR (= -2 log λ)[127], ne calcule pas de probabilités exactes, mais des approximations fiables des probabilités exactes[128]. Elle a une distribution χ² asymptotique (ou approximative). Les résultats de la mesure statistique du rapport de vraisemblance (G² ou LLR) sont facilement interprétables en tant que degrés d'association. Plus la valeur de LLR est élevée, plus la cooccurrence des deux mots est forte et statistiquement significative.

5.1.3.4 Conclusion

Les deux mesures d'association qui se prêtent le mieux à tout type d'associations, même aux moins fréquentes, sont la mesure statistique du rapport de vraisemblance (G² ou LLR) et le test du Fisher Exact (calcul hypergéométrique) (Weber, Vos & Baayen 2000 ; Evert & Krenn 2001 ; Evert 2002 ; Manning & Schütze 2002). En effet, Dunning (1933) montre que le test du chi-carré (χ²) a tendance à surestimer le degré d'association de mots rares, tout comme l'information mutuelle MI et la mesure du score Z, qui souffre également du problème de la distribution normale sous-jacente. Le test t en revanche semble générer de bons résultats (Evert & Krenn 2001).

Pour des corpus peu volumineux, le test du Fisher Exact, avec ses probabilités exactes, donne les meilleurs résultats (Evert 2002). Pour des corpus plus

---

[127] $G^2 = LLR = 2\ [\log L(k_1, n_1, p_1) + \log L(k_2, n_2, p_2) - \log L(k_1, n_1, p) - \log L(k_2, n_2, p)]$ (Dunning 1993)

avec $L(k, n, p) = p^k (1\text{-p})^{n-k}$

et donc $G^2 = 2\ [(k_1 * \log(p_1) + (n_1 - k_1) * \log(1 - p_1)) + (k_2 * \log(p_2) + (n_2 - k_2) * \log(1 - p_2)) - (k_1 * \log(p) + (n_1 - k_1) * \log(1 - p)) - (k_2 * \log(p) + (n_2 - k_2) * \log(1 - p))]$

ou en termes de fréquences observées et attendues :

$$G^2 = -2 \log \frac{L(O_{11}, C_1, r) \cdot L(O_{12}, C_2, r)}{L(O_{11}, C_1, r_1) \cdot L(O_{12}, C_2, r_2)}$$

avec $L(k, n, r) = r^k (1\text{-r})^{n-k}$

et avec $r = R_1/N$, $r_1 = O_{11}/C_1$, $r_2 = O_{12}/C_2$

(http://www.collocations.de/AM/index.html).

[128] Pour les explications détaillées sur la mesure du LLR : voir le chapitre 4 (Cf. 4.1.2).

volumineux, avec une distribution très déséquilibrée des fréquences, il vaut mieux recourir à la mesure statistique du rapport de vraisemblance, qui permet la comparaison des fréquences les plus faibles et des fréquences les plus importantes. Finalement, il convient de signaler que la mesure statistique du rapport de vraisemblance fournit les meilleurs résultats dans une fenêtre de 7 mots autour du mot de base alors que le test de Fisher Exact est le plus efficace dans une fenêtre de 5 mots (Weber, Vos & Baayen 2000).

5.2 LES COOCCURRENCES DES COOCCURRENCES

Afin de déterminer le degré de monosémie des spécificités, nous nous proposons d'aller au-delà du niveau des cooccurrences. En effet, nous visons à étudier les cooccurrences de deuxième ordre, c'est-à-dire les cooccurrences des cooccurrences d'un mot de base. Dans la première section de cette partie (5.2.1), nous ferons un bref survol des études ayant eu recours aux cooccurrences des cooccurrences et nous expliquerons l'intérêt de l'analyse. La deuxième section (5.2.2) sera consacrée au recoupement formel des cooccurrences des cooccurrences, qui constitue le point de départ de l'élaboration de la mesure de recoupement ou mesure de monosémie, expliquée dans la dernière partie de ce chapitre (Cf. 5.3).

5.2.1 Pourquoi les cooccurrences des cooccurrences ?

5.2.1.1 Les cooccurrences regroupées et interconnectées

Dans Véronis (2003 et 2004b), les cooccurrences d'un mot figurant dans un grand corpus sont regroupées suivant leur similarité ou dissimilarité (en fonction de leur co-fréquence) pour identifier les différents usages ou sens du « mot-cible » (ou mot de base). Concrètement, les cooccurrents les plus fréquents d'un mot-cible polysémique, qui n'apparaissent pas au contact les uns des autres, sont considérés comme des « mots-racines ». Ainsi, pour le mot-cible *barrage*, les mots-racines sont entre autres *eau* et *match*. Les autres cooccurrents du mot-cible sont voisins d'un de ces mots-racines, par exemple *ouvrage, rivière, cours* pour le mot-racine *eau*. Les cooccurrences de chaque mot-racine sont fortement interconnectées. Elles se caractérisent par une co-fréquence élevée et dès lors par une similarité sémantique importante. En outre, ces interconnexions permettent d'identifier et d'isoler « des composantes de forte densité » (Véronis 2003 : 268).

L'hypothèse avancée est que les différents usages ou sens d'un mot polysémique correspondent à ces composantes de cooccurrences interconnectées et

sémantiquement similaires[129]. Il est à noter que les cooccurrences sont identifiées dans la version lemmatisée du corpus, après suppression des mots-outils et des mots généraux. Pour procéder à l'analyse sémantique d'un mot de base, Véronis (2003 et 2004b) fait donc appel aux cooccurrences les plus fréquentes du mot de base. L'apport sémantique de ces cooccurrences (mots-racines) est précisé et enrichi par les autres cooccurrences du mot de base, qui apparaissent au contact de ces premières cooccurrences avec lesquelles elles sont fortement interconnectées.

L'approche, analogue, de Ji et al. (2003) s'appuie également sur les cooccurrences, appelés « contexonymes » (*contexonyms*), c'est-à-dire mots liés ou apparentés contextuellement (Ji et al. 2003). Les mots liés contextuellement s'avèrent en effet des indicateurs précieux du sens du mot de base dans un contexte donné. Le recours aux contexonymes permet de formaliser la relation de contexte entre les mots. Ces contexonymes sont situés dans un espace multidimensionnel par une méthode de classification hiérarchique, au moyen de « cliques » (sens minimaux des mots), ce qui permet de regrouper et de qualifier sémantiquement les contexonymes. Le but de cette approche par contexonymes est de repérer des associations inter-mots (cooccurrences et collocations) et des associations intra-mot (distinctions de sens du même mot), telles que les usages contextuels (*write a diary* versus *write an article*) ou les sens distincts de mots homonymiques ou polysémiques.

Dans notre étude, nous envisageons également de préciser et de qualifier sémantiquement les cooccurrences d'un mot de base, non pas en recourant aux autres cooccurrences (interconnectées) du mot de base, mais en faisant appel aux cooccurrences de ces cooccurrences.

L'approche contextuelle qui consiste à étudier les cooccurrences ou les contextes pour appréhender le sens d'un mot s'inscrit dans le cadre de la sémantique distributionnelle. Selon la sémantique distributionnelle, les écarts de sens se caractérisent par « une variation des contextes où figure un mot d'une partie à l'autre d'un corpus » (Habert et al. 2004 : 567). Habert et al. cherchent à détecter les mots

[129] Afin de calculer les sens d'adjectifs polysémiques, Venant (2004) recourt aussi au principe de graphes et de zones de forte densité. Cependant, ce ne sont pas des graphes de coooccurences, reposant sur des relations syntagmatiques (Véronis 2003 et 2004), mais des graphes de synonymes, qui s'appuient sur des relations paradigmatiques. Ainsi, pour un adjectif polysémique, le graphe de synonymes consiste en plusieurs sous-graphes ou cliques, qui correspondent à « une nuance possible de sens » pour l'adjectif (Venant 2004 : 1148). Pour la désambiguïsation des verbes, Jacquet et Venant (2005) recourent à ce même principe de graphes de synonymie, mais remplacent les noms propres ou les mots rares par leurs classes contextuelles. Par exemple, le mot *luth* est remplacé par la classe des instruments de musique dans le contexte « jouer du ».

qui ont plusieurs sens ou qui sont employés simultanément avec des sens divergents dans différentes parties du corpus. Afin de détecter ces mots aux sens mouvants, ils proposent de recourir aux cooccurrences de ces mots, étant donné qu'ils changent souvent de voisins. Ils avancent l'hypothèse qu'un mot à sens multiple « aurait des voisins moins proches entre eux qu'un mot univoque » (Habert et al. 2004 : 570).

Il s'ensuit que les mots homonymiques, polysémiques et vagues auraient des cooccurrences sémantiquement plus hétérogènes. En effet, les homonymes ont des contextes d'emploi souvent très différenciés. Toutefois, les sens des mots polysémiques, sémantiquement apparentés, ont plus de chances de se retrouver « dans des contextes proches » (Habert et al. 2004 : 566). Dès lors, les cooccurrences des mots polysémiques seront moins hétérogènes que celles des mots homonymiques. Ces observations confirment notre intention d'adopter l'idée d'un continuum d'homogénéité sémantique comprenant plusieurs degrés, en fonction des cooccurrences plus ou moins homogènes. Toutefois, en sémantique distributionnelle et contextuelle, deux problèmes majeurs se posent. D'une part, la distribution des différents sens d'un mot dans le corpus est souvent irrégulière et, d'autre part, « la répartition des traits permettant de classer les mots est souvent très éparpillée » (Habert et al. 2004 : 573). Pour remédier à ces problèmes, Habert et al. (2004) suggèrent de recourir aux cooccurrences et aux similarités de deuxième ordre.

De plus, les cooccurrences de premier ordre sont généralement des cooccurrences syntagmatiques du mot de base et parfois des cooccurrences paradigmatiques. Par contre, les cooccurrences de deuxième ordre ou d'ordre supérieur se caractérisent principalement par des relations paradigmatiques avec le mot de base (hyponymes, hyperonymes, synonymes, antonymes) (Pezik 2005) et dès lors, ces dernières sont plus intéressantes pour caractériser sémantiquement le mot de base.

5.2.1.2 Les cooccurrences des cooccurrences

Grefenstette (1994) propose de même des techniques de premier, de deuxième et même de troisième ordre afin de regrouper les mots et de découvrir des similarités sémantiques. Si les techniques de premier ordre étudient le contexte local, c'est-à-dire les cooccurrences autour du mot en question, les techniques de second ordre comparent les contextes du mot afin de découvrir des mots similaires. Les techniques de troisième ordre vont encore plus loin, en comparant des listes de mots similaires afin de les regrouper selon des axes sémantiques (Grefenstette 1994). Les mots qui partagent des « affinités de second ordre » (Grefenstette 1994 : 280), à savoir les mots presque synonymes et les mots apparentés sémantiquement, ne doivent pas nécessairement apparaître ensemble, mais ils se caractérisent par des contextes similaires.

De la même façon, les occurrences d'un mot (potentiellement ambigu) sont sémantiquement similaires si elles partagent des cooccurrences de deuxième ordre. Les cooccurrences de deuxième ordre permettent donc de vérifier si les cooccurrences (de premier ordre) sont sémantiquement homogènes ou non (Cf. 5.3).

5.2.1.3 Les cooccurrences des cooccurrences : la détection de synonymes

Les cooccurrences de deuxième ordre ou les cooccurrences des cooccurrences permettent entre autres de mettre en évidence des relations de synonymie (Martinez 2000). Pour un mot de base (ou un « pôle ») tel que *mesures*, Martinez (2000) calcule d'abord tous les cooccurrents de *mesures*, comme *nouvelles, unilatérales, concrètes, adopter, prises*. L'étape suivante consiste à calculer les cooccurrents des cooccurrents les plus spécifiques[130] (*nouvelles*, *prises*), ce qui revient à déterminer les cooccurrents de deuxième ordre ou les cooccurrents des cooccurrents, par exemple *décisions, dispositions, initiatives, monétaires, mesure*. Comme Martinez cherche les synonymes du pôle initial (*mesures*), il retient uniquement des cooccurrents de deuxième ordre qui apparaissent à la fois avec *nouvelles* et avec *prises* (Cf. figure 5.1).

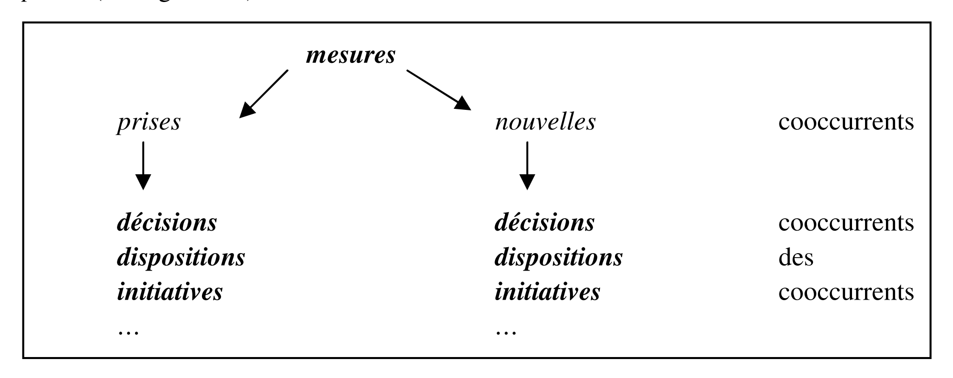


Figure 5.1 Cooccurrents des cooccurrents pour la détection de synonymes

Tant les cooccurrents que les cooccurrents des cooccurrents sont identifiés au niveau des formes fléchies, ce qui permet entre autres de préserver la distinction entre le singulier et le pluriel. Deux fenêtres d'observation sont utilisées, de 10 et de 20 mots respectivement autour du mot de base. La réitération du calcul permet non seulement de faire émerger des similarités distributionnelles d'un mot de base, mais également de trouver ses synonymes dans le corpus.

---

[130] Le logiciel Lexico3 permet d'indiquer la spécificité des cooccurrents pendant l'analyse automatique des cooccurrences (formule du calcul hypergéométrique (Lafon 1984)).

L'analyse de l'axe syntagmatique, effectuée à deux reprises (pour les cooccurrences et pour les cooccurrences des cooccurrences), contribue ainsi à la découverte de l'axe paradigmatique (les synonymes). Il est clair que les différents synonymes d'un mot de base sont des indices sémantiques précieux dans la perspective d'une mesure sémantique de monosémie (ou de polysémie).

5.2.1.4 Les cooccurrences d'ordre supérieur

D'après Denhière & Lemaire (2003), les cooccurrences de deuxième ordre et même d'ordre supérieur déterminent le degré d'association de deux mots M1 et M2, même si ces deux mots n'apparaissent jamais ensemble. Si les cooccurrences M1-M3 et M2-M3 sont suffisamment fortes, donc si leur degré d'association est suffisamment élevé, on considère que M1 et M2 sont associés et qu'ils sont des cooccurrents de deuxième ordre. Il est également possible d'extraire automatiquement les sens des mots à partir d'un réseau de cooccurrences lexicales de deuxième ordre, comme l'explique Ferret (2004). La connectivité des cooccurrents qui forment un sens est plus importante que leur connectivité avec les autres cooccurrents qui définissent les autres sens de ce mot (Ferret 2004).

Les travaux de Karov et Edelman (1998), qui se situent dans un contexte de désambiguïsation sémantique (ou WSD) et visent surtout à remédier au problème de rareté des données, recourent aux cooccurrences de deuxième ordre et d'ordre supérieur pour calculer la similarité de mots et de contextes. Celle-ci est définie en termes d'« usages similaires », car des mots similaires figurent dans des contextes similaires et leur proximité textuelle par rapport à un mot ambigu donne une indication du sens de ce mot. Des mots similaires se caractérisent également par des cooccurrents similaires de deuxième ordre, sans pour autant qu'ils aient les mêmes cooccurrents. Cette idée d'itération en matière de cooccurrents mène à une mesure de similarité transitive, qui prévoit une pondération en fonction de plusieurs critères (Cf. 5.1.2.6).

Les autres études et algorithmes recourant aux cooccurrences de deuxième ordre et d'ordre supérieur (Schütze 1998 ; De Marneffe & Dupont 2004) relèvent de l'acquisition sémantique. Elles se caractérisent par une approche vectorielle et matricielle et/ou par une décomposition en valeurs singulières (SVD)[131].

---

[131] SVD = Singular Value Decomposition (Cf. 5.1.1.2 pour une explication détaillée).

5.2.2 Le recoupement des cooccurrences des cooccurrences

Les cooccurrences des cooccurrences se révèlent particulièrement intéressantes pour déterminer le degré de (dis)similarité sémantique des cooccurrences et dès lors des occurrences d'un mot de base. Plus les cooccurrences sont spécifiques du mot de base et plus leur degré d'association avec le mot de base est fort, plus elles sont sémantiquement pertinentes. A ce sujet, Van Campenhoudt (2002b)[132] signale que les bigrammes[133] avec le degré d'association le plus élevé se constituent surtout de « noms propres, expressions figées empruntées à des langues étrangères, composés et cooccurrents appartenant à une langue technique ou de spécialité » (Van Campenhoudt 2002b : 21). Les bigrammes avec un degré d'association moyen sont des « mots de tous les jours » comme *telefonata anonima* (*coup de téléphone anonyme*) (Van Campenhoudt 2002b : 21). Les bigrammes avec un degré d'association plus faible « reflètent des structures syntaxiques et grammaticales » (Van Campenhoudt 2002b : 21).

Par conséquent, il serait plus judicieux de tenir compte de la significativité statistique des cooccurrences de premier et de deuxième ordre, afin de ne prendre en considération que les cooccurrences sémantiquement pertinentes. Ainsi, le degré de recoupement ou le degré d'homogénéité sémantique du mot de base serait calculé uniquement en fonction des cooccurrences les plus significatives statistiquement, c'est-à-dire les plus saillantes et donc, sémantiquement les plus pertinentes.

5.2.2.1 La saillance ou la significativité statistique des cooccurrences

Il est clair que les études visant à évaluer l'importance des caractéristiques des cooccurrences (à savoir les caractéristiques sémantiques, syntaxiques, position, etc.) intègrent toutes les cooccurrences du mot de base, même celles qui ne sont pas pertinentes (statistiquement significatives). Ces études n'adoptent donc aucune mesure d'association (de Loupy et al. 2000 ; Audibert 2003), même si certaines études d'évaluation des caractéristiques des cooccurrences recourent à la mesure de l'information mutuelle normalisée (Ferret 2004).

Les études faisant intervenir les cooccurrences de deuxième ordre ou d'ordre supérieur n'adoptent pas de mesure non plus, mais recourent à la décomposition en valeurs singulières pour l'identification des informations pertinentes et des proximités sémantiques (Schütze 1998 ; Denhière & Lemaire 20003 ; De Marneffe

---

[132] Van Campenhoudt (2002b) reprend la classification de Bindi et al. (1994), basée sur une étude de corpus.

[133] Un bigramme est une association de deux mots simples.

& Dupont 2004). Dans certaines études, la saillance ou la pertinence des collocations s'appuie principalement sur des critères de fréquence (Stevenson & Wilks 2001), sur la probabilité de co-fréquence et sur les autres éléments de la table de contingence (Véronis 2003 et 2004b) ou sur la co-fréquence et l'indice de Jaccard[134] (Habert et al. 2004 et 2005). D'autres études recourent à la probabilité exacte du calcul hypergéométrique (Heiden 2004 ; Martinez 2000).

Finalement, un certain nombre de recherches font état de la mesure d'association du rapport de vraisemblance (LLR) (Karov & Edelman 1998 ; Lapata 2002) ou d'une mesure analogue élaborée par Quasthoff & Wolff (Wandmacher 2005). En effet, les collocations ou les cooccurrences les plus indicatives d'un patron sémantique du mot de base se caractérisent par la valeur de LLR la plus élevée (Yarowsky 1994). Le LLR étant une mesure d'association stable et fiable (Cf. chapitre 4), nous préférons adopter cette mesure d'association également pour le calcul des cooccurrences et des cooccurrences des cooccurrences statistiquement significatives (Cf. 5.3). Le fait de recourir à la même mesure statistique, tant pour le calcul des spécificités que pour le calcul des cooccurrences (des cooccurrences), permet en outre de veiller à la cohérence méthodologique de notre étude.

Jusqu'à présent, la méthode des cooccurrences, y compris l'analyse des cooccurrences des cooccurrences, a été adoptée principalement dans des études de désambiguïsation sémantique et de recherche de synonymes ou de similarités sémantiques. Dans notre étude, nous nous proposons de recourir aux cooccurrences des cooccurrences dans un contexte de sémantique quantitative, plus particulièrement dans le but de mesurer le degré de recoupement ou le degré de monosémie des spécificités de notre corpus technique. Le degré de monosémie nous permettra de situer les spécificités dans un continuum de monosémie, allant des plus monosémiques aux moins monosémiques.

L'idée des degrés de monosémie ou degrés de polysémie est proposée également par Nerlich et al. (2003), où elle est exprimée en termes de « polysémie graduée ». Dans leur théorie graduée de la polysémie, Nerlich et al. (2003) relèvent des patrons sémantiques flexibles et avancent l'hypothèse que chaque mot est plus ou moins polysémique, avec des sens liés à un prototype par un ensemble de principes relationnels sémantiques, plus ou moins flexibles.

---

[134] L'indice de Jaccard est basé sur des proportions de fréquences observées et marginales (Cf. 5.1.3.1).

5.2.2.2 Homogénéité et hétérogénéité sémantique

En préparation de notre analyse sémantique automatisée et quantitative, nous avons d'abord mené une expérimentation sur un petit échantillon de 30 termes techniques, comprenant entre autres *broche, découpe, tour, avance*. Les dictionnaires techniques spécialisés et l'étude du contexte linguistique à partir des concordances nous ont permis, lors de cette analyse sémantique manuelle, d'accéder au(x) sens de ces 30 termes et de constater leur polysémie au sein du corpus technique. Par exemple, *broche* signifie (1) « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » et (2) « outil servant à usiner des pièces métalliques », *découpe* signifie (1) « action de découper » et (2) « résultat de la découpe (= pièce découpée) » et pour *tour*, nous recensons notamment les sens (1) « machine-outil pour l'usinage des pièces » et (2) « rotation ». Toutefois, pour l'analyse sémantique de 5000 spécificités du corpus technique, l'automatisation et la quantification s'imposent.

Dans le but d'opérationnaliser et de quantifier la monosémie, nous proposons d'implémenter la monosémie en termes d'homogénéité sémantique. Rappelons que les cooccurrences des cooccurrences permettent de vérifier dans quelle mesure le mot de base est monosémique ou homogène sémantiquement, parce que le degré de recoupement des cooccurrences de deuxième ordre est un indice important du degré de monosémie du mot de base (Cf. chapitre 2). Pour étudier le caractère monosémique ou polysémique d'une unité linguistique, on vérifie généralement si les contextes peuvent être considérés comme sémantiquement homogènes ou non (Condamines & Rebeyrolles, 1997). L'accès à la sémantique des cooccurrences pourra se faire (automatiquement) par le biais des cooccurrences de deuxième ordre. En effet, le degré de recoupement des cooccurrences de deuxième ordre indiquera si les cooccurrences de premier ordre (contextes du mot de base) sont similaires entre elles et si elles appartiennent au même champ sémantique (Cf. chapitre 2).

5.2.2.3 Homogénéité sémantique et monosémie traditionnelle

Afin de développer un critère d'analyse opérationnalisable et mesurable, nous proposons de recourir à cette mesure de monosémie ou de recoupement et d'implémenter la monosémie comme homogénéité sémantique. Par conséquent, les résultats de notre étude sémantique quantitative (Cf. chapitres 7 et 8) devront être interprétés et expliqués en fonction des choix méthodologiques de la mesure de monosémie élaborée. Il faudra en outre étudier les points de convergence et de divergence entre notre mesure de monosémie et ce que les monosémistes traditionnels considèrent comme monosémie ou polysémie. D'ailleurs, il est à noter qu'il n'est pas toujours très clair ce que les monosémistes traditionnels entendent par « monosémie ». En plus, il convient de signaler d'ores et déjà ce que notre mesure de monosémie permet de mesurer, mais aussi ce qu'elle ne permet pas de mesurer.

Dans un souci de précision et d'efficacité, les monosémistes de l'approche traditionnelle onomasiologique et prescriptive préconisent la monoréférentialité (chaque terme a un seul référent) et la monosémie (chaque terme a un seul sens) (Cf. chapitre 1). Ce sens unique est généralement prescrit par des ouvrages normatifs et expliqué (et / ou délimité) à l'aide d'une définition dans des normes ou dans un dictionnaire spécialisé. Néanmoins, il n'est pas toujours clair si ce sens prescrit s'applique effectivement à tous les contextes d'usage de l'unité terminologique.

Le fait d'implémenter la monosémie comme homogénéité sémantique permet certainement d'élaborer un continuum sémantique, allant de la plus grande homogénéité à la plus grande hétérogénéité sémantique, avec de nombreuses distinctions de degré entre ces deux extrémités. Les degrés d'homogénéité sémantique (ou de monosémie) et le continuum qui en résulte, conduisent à une analyse de régression simple qui étudie la corrélation entre le continuum de spécificité et le continuum de monosémie. Cependant, admettons d'emblée que l'hétérogénéité sémantique des cooccurrences des cooccurrences ne permet pas d'opérer une distinction tranchée entre l'homonymie, la polysémie et le vague, parce que les trois phénomènes se caractérisent par des cooccurrences sémantiquement hétérogènes, bien que ce soit à différents degrés.

Cette conséquence ainsi que les autres répercussions méthodologiques de notre mesure de monosémie seront expliquées dans la dernière partie (Cf. 5.3). En dépit de la lacune mentionnée ci-dessus, corollaire méthodologique de notre mesure de monosémie, nous tenons à insister sur son caractère innovateur. Elle permet non seulement de quantifier l'analyse sémantique, mais également d'opérationnaliser la monosémie en termes de degrés. La sémantique ainsi quantifiée et mesurée ne relève pas du discret, mais du continu et aboutit à l'établissement d'un continuum d'homogénéité sémantique.

5.3 MESURE DE RECOUPEMENT DES COOCCURRENCES DES COOCCURRENCES

Comme nous venons d'indiquer, nous tentons d'établir un continuum en quantifiant la monosémie et en opérationnalisant les critères de l'analyse sémantique par le recours à l'homogénéité sémantique. Comme un degré élevé d'homogénéité sémantique correspond à un degré élevé de recoupement des cooccurrences des cooccurrences (Cf. 5.2.2.2), le calcul du recoupement permettra de quantifier l'homogénéité sémantique, et donc la monosémie.

La première section de cette partie (5.3.1) décrira la préparation de la mesure de recoupement, à partir des cooccurrents et des cooccurrents des cooccurrents (5.3.1.1), le poids des cooccurrents des cooccurrents (5.3.1.2) et la formule pour la mesure de recoupement qui en découle (5.3.1.3). Dans la deuxième section (5.3.2), nous décrirons les différentes étapes de l'implémentation de la mesure de recoupement à l'aide d'un script en Python.

5.3.1 La préparation de la mesure de recoupement

5.3.1.1 Les cooccurrents et les cooccurrents des cooccurrents

Rappelons que si les cooccurrents de deuxième ordre se recoupent beaucoup, les cooccurrents de premier ordre seront sémantiquement plus homogènes ou plus similaires et indiquent un degré de monosémie plus élevé du mot de base. Les mots de base faisant l'objet du calcul du recoupement sont les quelque 5000 spécificités du corpus technique (Cf. chapitre 4).

Il est à noter que les cooccurrents de premier ordre (ou c), c'est-à-dire les cooccurrents directs du mot de base, seront considérés au niveau des types (*types*) : nous dresserons la liste de tous les cooccurrents différents, statistiquement significatifs, d'un mot de base (Cf. tableau 5.3). Leurs cooccurrents, c'est-à-dire les cooccurrents de deuxième ordre (ou cc), seront considérés par contre au niveau des occurrences (*tokens*), parce que nous prendrons en considérons tous les cooccurrents des cooccurrents statistiquement significatifs. Certains cooccurrents de deuxième ordre figureront probablement plusieurs fois dans la liste des cc, étant donné qu'ils apparaissent avec plusieurs cooccurrents (différents) de premier ordre (Cf. tableau 5.4). De telle façon, nous pourrons calculer le degré auquel les cooccurrents des cooccurrents se recoupent, c'est-à-dire la mesure dans laquelle les cooccurrents de deuxième ordre sont partagés par les cooccurrents de premier ordre.

Nous donnons ci-dessous un exemple :

mot de base (=spécificité)	c = cooccurrents de premier ordre (*types*)	cc = cooccurrents de deuxième ordre (*tokens*)
tour	*vertical*	*fraiseuse* *axes* *horizontal* *position* *tour*[135] …
	numérique	*fraiseuse* *axes* *commande* *perceuse* …
	minute	*heures* *secondes* *prend* …
	…	… …

Tableau 5.3 Mot de base + cooccurrents + cooccurrents des cooccurrents

Pour chaque mot de base, tous les cooccurrents statistiquement significatifs seront repérés. A cet effet, nous recourrons à la mesure d'association du rapport de vraisemblance (G^2 ou LLR), basée notamment sur la co-fréquence du mot de base et son cooccurrent. Nous proposons de respecter un seuil de significativité très sévère (à savoir une valeur $p < 0,0001$), afin de relever uniquement les cooccurrents les plus fortement associés et sémantiquement les plus pertinents. Par conséquent, dans la deuxième colonne des cooccurrents (c), chaque item figure une fois (*types*) (Cf. tableau 5.3).

Ensuite, la réitération du calcul des associations pour chaque cooccurrent (c) comme base (*node*) permet de repérer, par cooccurrent, tous ses cooccurrents statistiquement significatifs ($p < 0,0001$). Ainsi, dans la troisième colonne des cooccurrents des cooccurrents, chaque item pourra figurer soit une fois (s'il apparaît ensemble avec

[135] L'apparition du mot de base parmi dans les cc est également prise en considération.

un c de la liste des c), soit plusieurs fois (s'il apparaît ensemble avec plusieurs c de la liste des c). Les cc ou les cooccurrents des cooccurrents seront donc pris en considération comme occurrences (*tokens*) (Cf. tableau 5.3).

5.3.1.2 Le poids des cooccurrents des cooccurrents

Afin d'élaborer la formule du recoupement des cooccurrents des cooccurrents, il faudra d'abord déterminer le poids de ces cc pour le recoupement global. Une représentation schématique (Cf. tableau 5.4) fait intervenir une base, ses 5 c différents (c_1, c_2, c_3, c_4 et c_5) et tous leurs cc (25 au total). Ce schéma permettra d'expliquer le poids ou l'importance de chaque cc pour le recoupement global.

mot de base (=spécificité)	c = cooccurrents de premier ordre (*types*)	cc = cooccurrents de deuxième ordre (*tokens*)
base	c_1	x y z_1 z_2 z_3
	c_2	x y z_4 z_5 z_6
	c_3	w v z_7 z_8 z_9
	c_4	w v z_{10} z_{11} z_{12}
	c_5	w z_{13} z_{14} z_{15} z_{16}

Tableau 5.4 Mot de base + c + cc : schéma

Un cc partagé par tous les c, figure 5 fois dans la liste des cc, constituée de 5 blocs de cc (un bloc par c). Le cc figurant 5 fois aura donc un poids maximal de 5/5 (il figure dans 5 blocs des 5). Il pourra tout au plus figurer 5 fois dans la liste des cc (donc comme cooccurrent des 5 cooccurrents). Dans l'exemple, le recoupement maximal par cc correspond à 5/5 (=1) (Cf. tableau 5.5).

Par contre, un cc qui figure dans un seul bloc est un cc isolé car il est cooccurrent d'un seul des c (par exemple le cc z_1 du c c_1) et n'est pas partagé par d'autres c. Comme il figure une fois dans la liste des cc, il aura un poids minimal de 1/5. Dans l'exemple, le recoupement minimal par cc correspond à 1/5 (=0,2) (Cf. tableau 5.5).

poids par cc	recoupement	cc = cooccurrents de deuxième ordre (*tokens*)
poids maximal de 5/5	maximal	cc figure 5 fois sur 5
poids minimal de 1/5	minimal	cc figure 1 fois sur 5 (z_1)
poids de 2/5	moins important	cc figure 2 fois sur 5 (x)
poids de 3/5	plus important	cc figure 3 fois sur 5 (w)

Tableau 5.5 Poids des cooccurrents des cooccurrents

De même, le poids de *x* (figurant 2 fois dans la liste des cc ou dans 2 blocs) équivaut à 2/5 et le poids de *w* (figurant 3 fois dans la liste des cc) équivaut à 3/5 (Cf. tableau 5.5). Ainsi, on pourra calculer facilement le poids de chaque cc dans la liste de tous les 25 cc (*tokens*). Le poids de chaque cc correspond au rapport entre la fréquence du cc dans la liste des cc et le nombre de c (*types*).

Pour connaître le recoupement global, calculé à partir du recoupement de tous les cc, on fera d'abord la somme des poids individuels (donc 25 réitérations du calcul précédent des fractions 2/5 ou 3/5) et ensuite, le total (la somme des 25 fractions) sera divisé par 25 (le nombre total de cc (*tokens*) dans la liste). En effet, chaque cc contribue pour 1/25 au recoupement global calculé pour le mot de base.

5.3.1.3 La formule pour la mesure de recoupement

La formule pour la mesure de recoupement (Cf. figure 5.2) est basée sur le recoupement formel des cooccurrents des cooccurrents et prend en considération :

1) la fréquence d'un cc dans la liste des cc (= nombre de c apparaissant avec ce cc)	fq *cc*	p.ex. 3 (*w*)
2) le nombre total de c	nbr total *c*	p.ex. 5
3) le nombre total de cc	nbr total *cc*	p.ex. 25
et totalisant pour le nombre total de cc	$\sum_{cc}$	p.ex. 25

Rappelons qu'un cc sera d'autant plus important pour le recoupement total, s'il figure plus souvent dans la liste des cc, c'est-à-dire si sa fréquence dans la liste des cc est plus élevée ou s'il est plus souvent partagé par les cooccurrents ou c.

$$\sum_{cc} \frac{\text{fq}\, cc}{\text{nbr total}\, c \cdot \text{nbr total}\, cc}$$

Figure 5.2 Mesure de recoupement

Le résultat de la formule se situe toujours entre 0 (hétérogénéité sémantique – pas de recoupement) et 1 (homogénéité sémantique – recoupement parfait). Plus le résultat s'approche de 1, plus le recoupement est important et plus les cc seront fortement partagés globalement. Un recoupement très important est, on l'a vu, une indication de l'homogénéité sémantique du mot de base. Plus le résultat s'approche de 0, plus il est faible, plus le recoupement est faible. Si les cc sont peu partagés globalement,

cela indique une distribution plus hétérogène des cooccurrents et dès lors moins d'homogénéité du mot de base.

Verbalisons, par souci de clarté, la formule de la mesure de recoupement et reprenons en guise d'exemple le cc fortement partagé (*w*) du schéma (Cf. tableau 5.4), partagé par 3 c des 5 c au total. Cela veut dire que 3 c des 5 c apparaissent avec ce cc en question, ce qui indique un recoupement plutôt important. Dans le numérateur de la formule, nous incluons le nombre de c qui ont ce cc en commun (fq *cc*), en l'occurrence 3, dans le dénominateur nous incluons le nombre total de c différents (au niveau des *types*), en l'occurrence 5. Le recoupement est donc exprimé par la fraction 3/5. En exprimant pour chaque cc le recoupement par la fraction *nombre de c avec le cc (*ou *fq cc)* divisé par *nombre total de c*, le résultat se situe toujours entre 0 (pas ou peu de recoupement) et 1 (recoupement important ou parfait) et par conséquent, le résultat est facilement interprétable. Comme on fait le total pour tous les cc, il faut ajouter dans le dénominateur le nombre total de cc (au niveau des *tokens*), car on considère en effet tous les cc (*tokens*) évidemment avec les doublons responsables du recoupement formel.

Soulignons, une fois de plus, qu'il ne s'agit pas du nombre de cc différents (*types*), mais du nombre total de cc (*tokens*), à savoir tous les mots (cc) qui cooccurrent avec tous les c différents (*types*) relevés pour le mot de base.

Une des conséquences du caractère novateur de notre mesure de recoupement des cooccurrences des cooccurrences est qu'il n'existe pas de mesure de référence ou de *Gold Standard* permettant d'évaluer les résultats quantitatifs de notre mesure. Nous proposons dès lors de procéder à une comparaison manuelle des cooccurrences les plus saillantes et les plus pertinentes d'un certain nombre de spécificités. Cette analyse permettra de vérifier si le degré de monosémie (ou d'homogénéité sémantique) calculé pour une spécificité (mot de base) est justifié par la (dis)similarité sémantique des cooccurrences sémantiquement pertinentes et statistiquement significatives.

5.3.2 La concrétisation de la mesure de recoupement

La concrétisation de ce deuxième axe méthodologique de l'analyse des cooccurrences consiste à appliquer la mesure de recoupement aux spécificités du corpus technique afin de calculer leur degré de monosémie. A cet effet, nous avons réalisé un algorithme à partir de scripts en Python. Cet algorithme consiste en plusieurs étapes, dont les détails sont précisés dans le document en annexe (Cf. annexe 8).

5.3.2.1 Les cooccurrents et les cooccurrents des cooccurrents

Les fichiers *.cnr[136] du corpus technique font l'objet de deux analyses de cooccurrences. D'abord, une première analyse prend la spécificité comme base (lemme). Pour tous les lemmes, elle répertorie tous leurs cooccurrents (formes graphiques), dans une fenêtre d'observation[137] de 5 mots à gauche et 5 mots à droite. Ensuite, une deuxième analyse prend le cooccurrent comme base et vise à repérer tous ses cooccurrents, donc les cooccurrents de deuxième ordre. Cette deuxième analyse des cooccurrences prendra ainsi comme base toutes les formes graphiques et répertorie tous leurs cooccurrents (formes graphiques), également dans une fenêtre d'observation de 5 mots à gauche et 5 mots à droite.

Les paramètres modifiables sont le type de cooccurrent à relever (lemme ou forme fléchie) et la fenêtre d'observation. Nous optons pour une fenêtre de [-5,+5], parce qu'elle apporte suffisamment d'informations sémantiques, sans qu'il y ait trop de bruit, et qu'elle permet un traitement informatique efficace.

Au premier niveau d'analyse de la spécificité, la base de la cooccurrence est nécessairement relevée sous forme lemmatisée, puisqu'il faut pouvoir rattacher les informations sémantiques (degré de monosémie) aux informations de spécificité (degré de spécificité) (Cf. chapitre 4). Le choix du lemme pour le mot de base repose donc sur des critères méthodologiques. Par ailleurs, pour le cooccurrent (ou le collocatif[138] de la combinaison de mots), la forme graphique ou forme fléchie s'impose, en raison des informations sémantiques plus riches qu'elle véhicule (Cf. la différence sémantique entre *pièce à usiner* et *pièce usinée*, par exemple).

---

[136] Rappelons qu'un fichier *.cnr du corpus technique est la version lemmatisée et catégorisée d'un fichier texte. Le fichier *.cnr se constitue de trois colonnes, à savoir (1) forme graphique, (2) lemme et (3) code Cordial (*POS-tag*). Les trois colonnes sont divisées par des tabulations, ce qui facilite la recherche de données et la programmation en Python.

[137] Signalons que la fenêtre d'observation actuelle [-5 ; +5] ne tient pas compte des frontières de documents. Dans un premier temps, toutes les formes graphiques cooccurrentes avec le mot de base sont intégrées dans la base de données. Ensuite, le seuil de significativité très sévère ($p < 0,0001$) permet de supprimer les cooccurrents non significatifs, notamment les cooccurrents erronés qui figurent dans le document suivant ou précédent. Ainsi, le seuil de significativité permet de limiter le bruit engendré par la transgression des frontières de documents. Etant donné que les fiches sont composées des documents les plus courts, ce problème se pose le plus dans ce sous-corpus. A cet effet, nous avons procédé à la génération d'une base de données de 2 mots à gauche et 2 mots à droite pour les fiches, ce qui permet de limiter le problème de la transgression des frontières de documents (Cf. chapitre 8).

[138] Collocatif (*collocate*) : le cooccurrent du mot de base (Cf. 5.1.2.1).

Puisque ce collocatif est la base du deuxième niveau d'analyse, la forme fléchie s'impose également à ce deuxième niveau tant pour la base que pour le collocatif. Ainsi, le choix de la forme graphique ou forme fléchie pour les cooccurrents et pour les cooccurrents des cooccurrents s'explique principalement par des raisons d'ordre sémantique.

5.3.2.2 Le calcul des degrés d'association

Les informations de cooccurrence (pour les 12 fichiers *.cnr) sont fusionnées et enregistrées sous forme de deux bases de données, une première pour les cooccurrences lemme – forme graphique (i.e. mot de base – cooccurrent) et une deuxième pour les cooccurrences forme graphique – forme graphique (c – cc) (Cf. annexe 8). Ces deux bases de données comprennent les données de cooccurrence suivantes : collocatif, base, co-fréquence, cfreq[139], nfreq[140].

Comme on dispose de tous les éléments requis de la table de contingence, le traitement statistique et la mesure d'association du rapport de vraisemblance (G^2 ou LLR) permettent d'obtenir deux bases de données avec les données statistiques suivantes : collocatif, base, co-fréquence, valeur de LLR (ou degré d'association), valeur p. Cette dernière permettra des opérations de sélection en fonction du seuil de significativité plus ou moins sévère. Les deux bases de données sont enfin fusionnées en une grande base de données à deux niveaux :

1) au niveau 1 : lemme (= spécificité) + forme graphique (= cooccurrents)

2) au niveau 2 : forme graphique comme mot de base (= cooccurrents du niveau précédent ou cooccurrents de premier ordre) + forme graphique (= cooccurrents des cooccurrents ou cooccurrents de deuxième ordre).

5.3.2.3 Le calcul des degrés d'homogénéité sémantique

Finalement, cette double base de données sera indexée et interrogée. L'indexation est une opération technique qui facilite les recherches en réduisant considérablement le temps de recherche du script en Python. Pour chaque spécificité, la base de données indexée sera interrogée afin de calculer le recoupement des cooccurrents des cooccurrents. A cet effet, la fonction Python de l'algorithme prévoit les paramètres suivants : la base (spécificité à analyser), le seuil de significativité pour

[139] La fréquence du collocatif avec n'importe quelle base.

[140] La fréquence de la base avec n'importe quel collocatif. Les fréquences 'cfreq' et 'nfreq' et la co-fréquence permettent de compléter la table de contingence (Cf. 5.1.3 tableau 5.1).

les cooccurrents de premier ordre (p.ex. 0,95 pour p < 0,05), le seuil pour les cooccurrents de deuxième ordre et, finalement, la base de données. Rappelons que nous préférons adopter un seuil de significativité très sévère (seuil de 0,9999 pour une valeur p < 0,0001), afin de relever uniquement les cooccurrents et les cooccurrents des cooccurrents sémantiquement pertinents et donc de quantifier le recoupement de ces derniers.

Il reste à signaler, du point de vue méthodologique, que le calcul du degré de recoupement ne pourra pas se faire pour les spécificités avec 0 c ou avec 1 c statistiquement significatif (Cf. chapitre 6). S'il n'y a pas de c, il n'y a pas de cc et dès lors, la formule s'avère inapplicable. S'il y a un seul c, le recoupement de ses cc est impossible et par conséquent le calcul du degré de recoupement n'a pas de sens. En plus, un hapax qui figure une fois dans le corpus pourra difficilement afficher plusieurs usages par le biais de son occurrence unique. Par conséquent, nous avons décidé de supprimer dans la liste des spécificités les hapax ainsi que les spécificités avec 0 c et 1 c (Cf. chapitre 6). Une fonction en Python permet de dénombrer le nombre de cooccurrents d'un mot de base (spécificité) et dès lors de supprimer les spécificités en fonction du nombre de cooccurrents au seuil de significativité choisi (0,9999). Il en résulte une liste de 4717 spécificités (Cf. annexe 7).

Pour les 4717 spécificités du corpus technique, nous pouvons ainsi calculer le degré de recoupement et donc le degré de monosémie (ou d'homogénéité sémantique) qui permettra de situer les spécificités sur un continuum de monosémie. Par définition, les mots avec un degré de monosémie identique auront le même rang de monosémie, par analogie avec le rang de spécificité.

Chapitre 6
Mises au point méthodologiques

Le sixième chapitre marque la transition entre les deux chapitres méthodologiques précédents (Cf. chapitres 4 et 5) et les deux chapitres qui présentent les résultats des analyses statistiques (Cf. chapitres 7 et 8). Avant de déterminer définitivement le degré de recoupement des spécificités et, dès lors, leur rang de monosémie, il convient de procéder à quelques mises au point méthodologiques. A cet effet, la mesure de recoupement élaborée dans le chapitre précédent, sera soumise à des expérimentations, permettant de déterminer la configuration la plus stable et de mieux comprendre la formule de la mesure de recoupement. Les premiers résultats exploratoires des expérimentations constituent ainsi la première étape du processus d'interprétation des résultats de l'analyse.

Les expérimentations feront l'objet de la première partie de ce chapitre (6.1). La deuxième partie (6.2) sera consacrée aux vérifications nécessaires pour mieux comprendre l'impact des différents facteurs dans le numérateur et le dénominateur de la formule. Finalement, dans la dernière partie (6.3), nous procéderons à l'élaboration d'une mesure de recoupement technique, en fonction de la spécificité ou technicité des cooccurrents des cooccurrents, dans le but de préciser et de nuancer les résultats de la mesure de recoupement de base.

6.1 LA CONFIGURATION IDÉALE

Les questions principales conduisant à la configuration idéale de la base de données portent sur trois paramètres, à savoir la forme graphique ou la forme canonique des cooccurrents (6.1.1), la taille de la fenêtre d'observation (6.1.2) et le seuil de significativité (6.1.3). Ces trois questions s'appliquent tant au niveau 1 des cooccurrents qu'au niveau 2 des cooccurrents des cooccurrents. Nous procéderons également à des analyses qui font varier simultanément plusieurs paramètres de configuration (6.1.4). La comparaison des résultats de plusieurs configurations alternatives permettra à la fois de déterminer la configuration la plus stable et de fournir les informations sémantiques les plus stables, tant en termes de degré de recoupement ou de monosémie qu'en termes de rang de monosémie.

6.1.1 La forme graphique ou la forme canonique ?

Le premier paramètre oppose la forme graphique (forme fléchie) à la forme canonique (lemme), pour les cooccurrents et pour les cooccurrents des cooccurrents. L'impact de ce paramètre est analysé pour un échantillon (Ntec02.cnr) d'environ 320.000 occurrences de la revue *Trametal*. Les expérimentations portent sur les 25 spécificités les plus spécifiques du corpus technique entier.

Le tableau ci-dessous (Cf. tableau 6.1) visualise les 25 spécificités numérotées et leur degré de recoupement (Cf. dernière colonne), au seuil de significativité de 0,9999, pour la configuration de base LWWtec02 (LWW = *Lemma – Wordform – Wordform*), donc lemme – forme fléchie – forme fléchie. Les spécificités se situent au niveau des lemmes, les cooccurrents et les cooccurrents des cooccurrents sont repérés au niveau des formes fléchies.

N° du degré de spécificité	spécificité	N° du degré de recoupement	spécificité	degré de recoupement
1	*machine*	9	***Fig***	0,1383
2	*outil*	5	***mm***	0,1045
3	*usinage*	18	*précision*	0,0973
4	*pièce*	21	*type*	0,0967
5	*mm*	16	*fraisage*	0,0863
6	*vitesse*	24	*gamme*	0,0804
7	*coupe*	6	*vitesse*	0,0793
8	*broche*	3	*usinage*	0,0784
9	*Fig*	13	*diamètre*	0,0768
10	*axe*	20	*surface*	0,0767
11	*copeau*	11	*copeau*	0,0761
12	*plaquette*	7	*coupe*	0,0761
13	*diamètre*	14	*commande*	0,0745
14	*commande*	17	*arête*	0,0742
15	*acier*	19	*usiner*	0,0734
16	*fraisage*	22	*système*	0,0709
17	*arête*	23	*fraise*	0,0708
18	*précision*	8	*broche*	0,0678
19	*usiner*	10	*axe*	0,0677
20	*surface*	25	*permettre*	0,0652
21	*type*	4	***pièce***	0,0652
22	*système*	12	*plaquette*	0,0609
23	*fraise*	15	*acier*	0,0566
24	*gamme*	1	***machine***	0,0505
25	*permettre*	2	***outil***	0,0481

Tableau 6.1 Les 25 spécificités et leur degré de recoupement dans LWWtec02

Les spécificités sont indiquées dans la deuxième colonne du tableau ci-dessus (Cf. tableau 6.1), leur degré de recoupement dans la cinquième, les colonnes 3 à 5 étant classées par ordre décroissant de degré de recoupement. Parmi les mots les plus monosémiques (en tête de liste dans la colonne 4), on trouve *Fig* et *mm*, intuitivement plus monosémiques en effet. Les mots les plus polysémiques se situent en bas de liste, tels que *machine*, *outil* et *pièce*. Les résultats ci-dessus (Cf. tableau 6.1) sont indicatifs, le degré de recoupement étant calculé sur un corpus restreint.

Dans les résultats du tableau 6.1, le degré de recoupement est déterminé et calculé pour les formes fléchies des c et des cc. Comme nous l'avons évoqué ci-dessus, les formes fléchies apportent des informations sémantiques plus riches et permettent de faire la distinction entre, par exemple, *pièce à usiner* et *pièce usinée*. Toutefois, la question se pose de savoir si la prise en considération du lemme des c ou du lemme des c et des cc influence le degré de recoupement et dès lors le classement des spécificités. Signalons d'emblée qu'au niveau du lemme, on recense moins de c et moins de cc différents (*types*), étant donné que les c et les cc seront regroupés sous leur lemme correspondant. Par conséquent, les cc (lemmes) pourraient manifester un degré de recoupement ou de monosémie plus élevé[141] et donc un degré d'homogénéité sémantique plus élevé pour le mot de base. Cependant, les différences de degré de recoupement ou de monosémie ne se traduisent pas toujours par des différences de rang de monosémie. Si la plupart des mots de base (spécificités) se caractérisent par un degré de recoupement plus élevé au niveau des lemmes, ils auront plus ou moins le même rang de monosémie pour les lemmes des cc que pour les formes fléchies des cc, parce que le rang de monosémie est accordé en fonction du tri des spécificités par ordre décroissant de degré de recoupement.

Afin de vérifier les différences de degré et de rang de monosémie en fonction de la forme graphique (fléchie) ou canonique (lemme) des c et des cc, trois bases de données sont générées au seuil de significativité de 0,9999, pour une fenêtre d'observation de 5 mots à gauche et 5 mots à droite (Cf. tableau 6.2). Rappelons que, pour l'analyse des spécificités, les spécificités sont toujours considérées au niveau des lemmes. Par conséquent, c'est le lemme qui s'impose pour les analyses de cooccurrences et qui permet de rattacher les informations de cooccurrence au mot spécifique en question (Cf. chapitre 5).

---

[141] Les formes fléchies des cc comprennent beaucoup de substantifs, tant au singulier qu'au pluriel, par exemple *débit – débits*, ainsi que les formes conjuguées des verbes (participe présent, participe passé). Toutefois, dans la base de données des lemmes des c et des cc, les cooccurrences pertinentes (au niveau des c et des cc) sont calculées en fonction des co-fréquences des lemmes, ce qui peut donner lieu à des différences considérables en matières de cc significatifs retenus.

Base de données	spécificités = mots de base	c = cooccurrents de premier ordre	cc = cooccurrents de deuxième ordre
LWWtec02	lemmes	formes fléchies	formes fléchies
LLWtec02	lemmes	lemmes	formes fléchies
LLLtec02	lemmes	lemmes	lemmes

Tableau 6.2 La configuration des bases de données LWW, LLW, LLL

En regardant les degrés de recoupement dans les trois configurations LWW, LLW et LLL (Cf. figure 6.1 ci-dessous), on observe que le degré de recoupement est généralement plus élevé lorsque les c et les c et cc sont des lemmes (Cf. *copeau, acier*). En effet, la différence entre LLW et LLL est plus grande que celle entre LWW et LLW. Les mots les plus hétérogènes sémantiquement (et les plus spécifiques) (*machine, outil*) ont un degré de recoupement plus faible dans les trois configurations. Les mots les plus homogènes sémantiquement (*Fig, mm*) ont un degré de recoupement plus élevé, généralement dans les trois configurations, à l'exception de *copeau*.

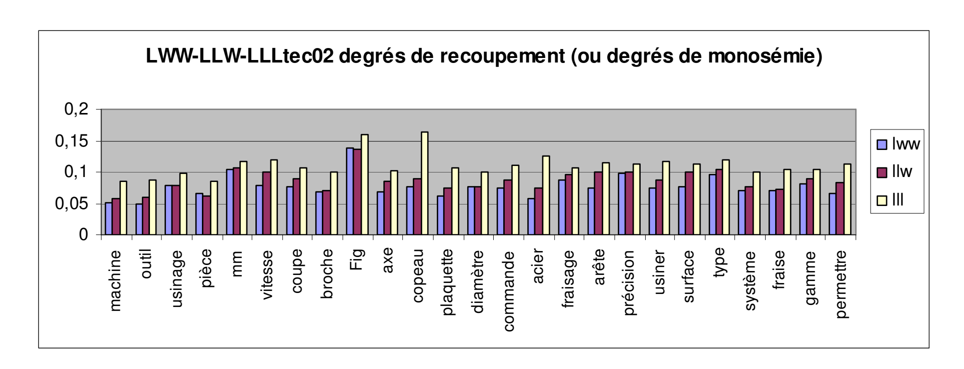


Figure 6.1 Degrés de recoupement dans LWWtec02, LLWtec02, LLLtec02

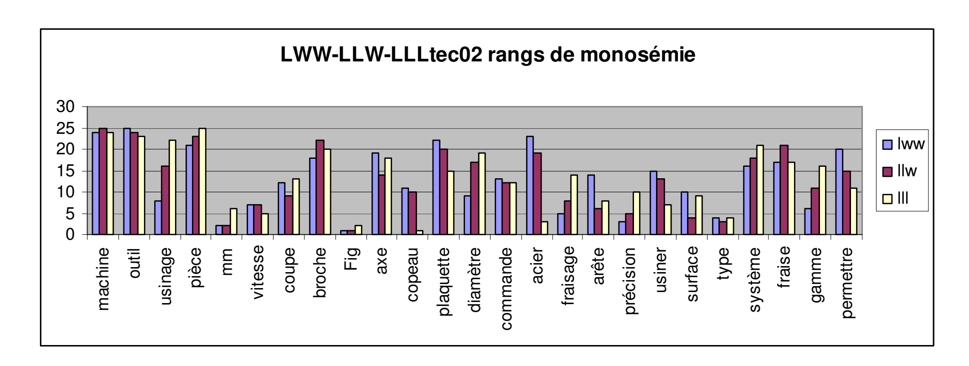


Figure 6.2 Rangs de monosémie dans LWWtec02, LLWtec02, LLLtec02

La figure 6.2 visualise les rangs de monosémie dans les trois configurations. Des rangs de monosémie de 1 ou 2 signifient que les mots en question sont les plus monosémiques, par exemple *Fig* et *mm*. En revanche, les rangs 24 et 25 caractérisent les mots les plus hétérogènes sémantiquement, en l'occurrence *machine* et *outil*. Les barres de cet histogramme groupé (Cf. figure 6.2) sont donc inversement proportionnelles aux barres de l'histogramme groupé précédent (Cf. figure 6.1). Il est clair que le rang de monosémie d'un mot change, non seulement en fonction de son propre degré de monosémie, qui est plus ou moins élevé dans les différentes configurations, mais également en fonction du rapport entre son propre degré de monosémie et le degré de monosémie des autres mots de la sélection. En effet, un mot avec un degré de recoupement similaire dans les trois configurations, pourra quand même se voir attribuer un rang de monosémie plus élevé (donc plus polysémique) si les autres mots de la sélection ont un degré de recoupement plus élevé et s'ils acquièrent, de ce fait, un rang de monosémie plus bas (plus près de 1, donc plus monosémique).

Pour ces trois configurations, à savoir LWW, LLW et LLL, les différences de rang les plus importantes s'observent pour *copeau* et *acier*. La spécificité *acier*, qui était plutôt hétérogène sémantiquement dans LWW, est très homogène sémantiquement dans LLL. Cela signifie que les lemmes des cc se recoupent beaucoup plus que les formes fléchies des cc. Si certains mots acquièrent un rang plus monosémique en passant des formes fléchies aux lemmes, d'autres se voient accorder, de ce fait, un rang plus polysémique, tels que *usinage*.

Les différences de degré et de rang évoquées ci-dessus sont certes indicatives pour certaines spécificités, mais ne permettent pas de visualiser les tendances globales. C'est la raison pour laquelle nous recourons à la technique de Multidimensional Scaling (MDS)[142] ou de positionnement multidimensionnel. Le MDS permet d'analyser une matrice de proximité (de similarité ou de dissimilarité) établie pour un ensemble de données. Le but est de modéliser les similarités ou dissimilarités entre les données à partir de leurs valeurs, afin de visualiser ces données dans un espace à deux dimensions. Cette technique vise surtout à réarranger les données de façon efficace, afin d'obtenir une configuration visuelle des distances observées. L'interprétation des dimensions et du positionnement se révèle parfois difficile, car il

[142] Le MDS est une méthode d'analyse multivariée descriptive, telle que l'analyse factorielle des correspondances (AFC) ou l'analyse en composantes principales (ACP). A la différence de ces techniques, le MDS permet d'analyser tout type de matrice de (dis)similarité, si les (dis)similarités sont évidentes. Le MDS n'impose pas de restrictions, telles que des relations linéaires entre les données sous-jacentes, leur distribution normale multivariée ou la matrice de corrélation (Cf. analyse factorielle) (http://www.statsoft.com/textbook/stmulsca.html).

est impossible d'interpréter les axes par la corrélation avec les variables analysées. Nous n'entrons pas dans les détails du calcul, parce que le Multidimensional Scaling est implémenté dans le logiciel d'analyse statistique R[143], qui permet une utilisation plus conviviale. Dans le logiciel R, il existe deux types de Multidimensional Scaling[144], à savoir le *isoMDS* et le *cmdscale*. Comme le isoMDS est un type de positionnement non métrique, il est plus flexible. Les deux types prennent comme point de départ une matrice de dissimilarité et indiquent la distance de chaque donnée (ou variable) par rapport aux autres. Cette matrice de dissimilarité est également générée par R. Le résultat du MDS est une visualisation (*plot*) selon deux axes, qui présentent les données analysées et leurs distances, tout comme les regroupements de données et les données isolées ou périphériques.

Nous recourons au MDS à des fins d'exploration et principalement dans le but de visualiser les rangs de monosémie des 25 spécificités dans les trois configurations (LWWtec02 – LWWtec02 – LLLtec02). A cet effet, un document (*.txt) avec les rangs de monosémie de ces 25 spécificités dans les trois configurations a été introduit dans le logiciel R (Cf. tableau 6.3). L'analyse de MDS permet de visualiser la distance entre les mots, représentés par leur rang de monosémie, en faisant varier les configurations (Cf. figure 6.3).

	LWW	LLW	LLL
machine	24	25	24
outil	25	24	23
usinage	8	16	22
…	…	…	…

Tableau 6.3 MDS des 25 spécificités

La visualisation ci-dessous du MDS pour les 25 spécificités et dans les trois configurations (Cf. figure 6.3) montre que les mots les plus polysémiques se regroupent et se situent à gauche de la visualisation (*machine, outil*). Les mots les plus monosémiques se regroupent également et se situent à droite (*Fig, mm, type*). Dans cette représentation visuelle du MDS, l'axe horizontal pourra donc s'interpréter comme l'axe sémantique, allant des mots plus polysémiques à gauche aux mots plus monosémiques à droite. L'axe vertical pourra s'interpréter comme l'axe de la stabilité dans les trois configurations. Les mots avec des différences de

---

143 Http://www.r-project.org.

144 Dans R : bibliothèque « MASS » (Cf. Venables, W. N. & B. D. Ripley 2002. *Modern Applied Statistics with S*. Fourth edition. Springer).

rang importantes dans les trois configurations se trouvent à une distance plus importante des autres mots. Ainsi, le mot *acier* se situe en haut de la visualisation, puisqu'il est plutôt polysémique dans la configuration LWW et monosémique dans LLL. Le mot *usinage* en revanche se situe en bas de la visualisation, étant donné qu'il se trouve parmi les mots les plus monosémiques dans la configuration LWW et qu'il est plutôt polysémique dans LLL. La plupart des mots se trouvent bien au milieu de la visualisation et se caractérisent par une relative stabilité de leur rang de monosémie dans les trois configurations.

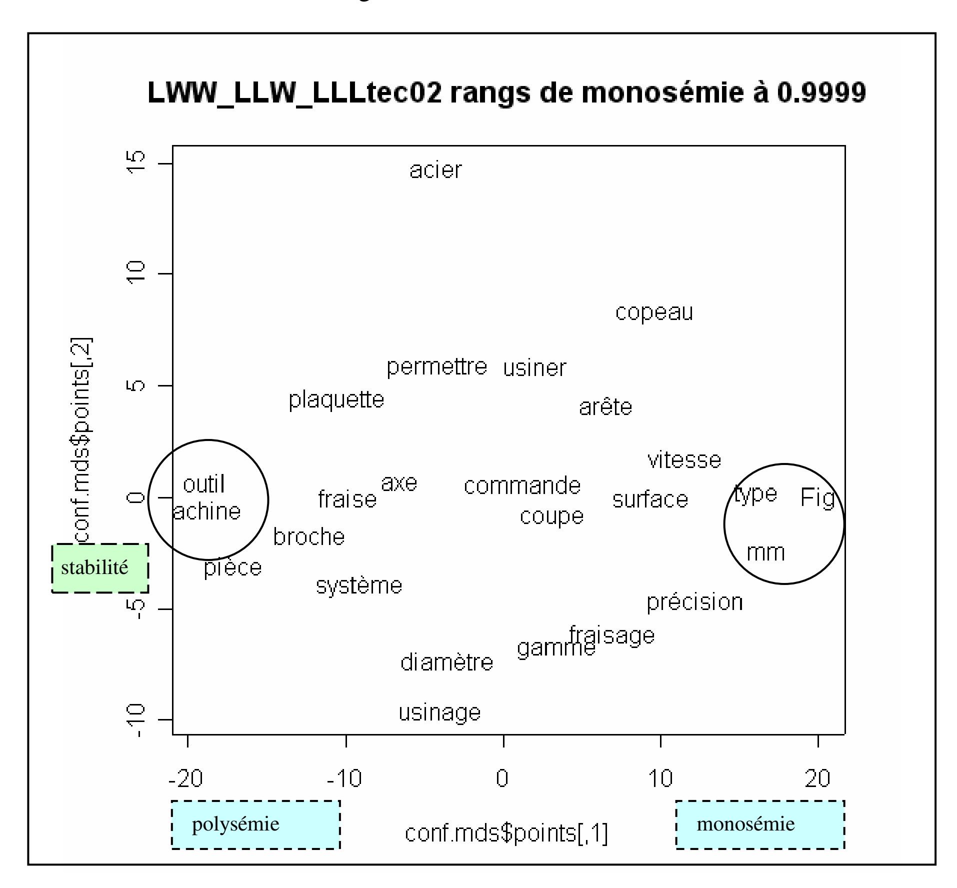


Figure 6.3 Résultat MDS des 25 spécificités (dans les trois configurations)

Le choix méthodologique qui consiste à identifier les c et les cc au niveau des formes fléchies se justifie donc par des résultats similaires dans les trois configurations et cela pour la plupart des 25 spécificités. Bien que le degré de recoupement soit généralement plus élevé pour les lemmes des cc, les différences de rang de monosémie ne sont pourtant pas spectaculaires.

Pour répondre à la question principale de notre étude, nous nous intéressons principalement au rang de monosémie des spécificités, plutôt qu'à leur degré de recoupement. Or, il est clair que des analyses plus approfondies s'imposent, qui font intervenir non seulement ces trois configurations, mais également les autres paramètres (Cf. 6.1.4), tels que la taille de la fenêtre d'observation et le seuil de significativité, que nous ferons d'abord varier séparément (Cf. 6.1.2 et 6.1.3).

6.1.2 La taille de la fenêtre d'observation

Afin de vérifier la taille idéale de la fenêtre d'observation (*span*), nous procédons à une comparaison de différentes tailles. Les expérimentations sont conduites également pour les 25 mots les plus spécifiques, sur le même échantillon (Ntec02.cnr) et au même seuil de significativité de 0,9999 (p < 0,0001). Les fenêtres d'observation comparées sont de taille 1, 2, 3, 4, 5, 6, 8, 10, 12, 15 et 3-15 (à partir du 3e mot à droite et à gauche jusqu'au 15e mot inclus). Cette dernière fenêtre d'observation est intéressante, car elle permet d'exclure les cooccurrents syntaxiques et de se concentrer surtout sur les cooccurrents lexicaux.

Les expérimentations ont pour but de vérifier si la taille de la fenêtre d'observation préconisée de 5 mots à gauche et 5 mots à droite du mot de base (ou [-5;+5]) n'est pas périphérique par rapport aux autres tailles. Une fenêtre plus petite entraîne certes moins de bruit, mais aussi moins de c (et cc) sémantiquement pertinents (notamment des collocations) et plus de c (et cc) syntaxiquement dépendants. Une fenêtre plus large apporte plus de c (et cc) sémantiquement pertinents, mais risque d'inclure plus (trop ?) de bruit. Nous étudierons l'impact du bruit lorsque la fenêtre est très large (> 10), ainsi que les patrons qui se dégagent à travers les tailles différentes.

Dans les 11 bases de données de tailles différentes (toujours au seuil de 0,9999), on détermine le rang de monosémie pour les 25 spécificités les plus spécifiques. Les rangs de monosémie sont enregistrés dans deux documents *.txt, premièrement avec les 25 spécificités comme rangées (Cf. tableau 6.4) et deuxièmement avec les 11 tailles de fenêtre d'observation (*span*) comme rangées (Cf. tableau 6.5). Nous procédons à une analyse de MDS, principalement pour vérifier les distances entre les différentes tailles (*span*) pour les rangs de monosémie des 25 mots les plus spécifiques (Cf. figure 6.6 plus loin).

	span1	span2	span3	…
machine	23	24	25	…
outil	25	25	24	…
usinage	21	17	11	…
…	…	…	…	…

Tableau 6.4 MDS des 25 spécificités

	machine	*outil*	*usinage*	…
span1	23	25	21	…
span2	24	25	17	…
span3	25	24	11	…
…	…	…	…	…

Tableau 6.5 MDS des 11 tailles différentes

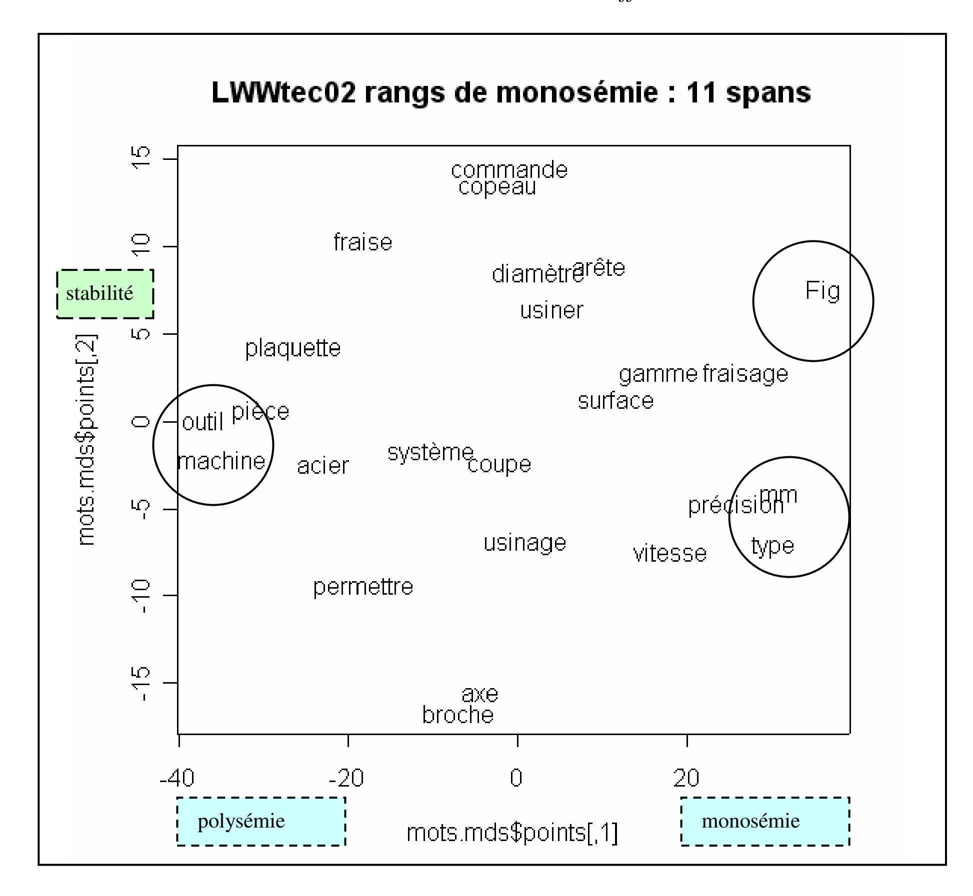


Figure 6.4 Résultat MDS des 25 spécificités (pour les 11 tailles)

La première visualisation des distances entre les 25 spécificités (Cf. figure 6.4), confirme les résultats précédents concernant les distances à travers les trois configurations (LWW, LLW, LLL) (Cf. figure 6.3). Il est clair que les mots les plus polysémiques se regroupent à gauche (*machine, outil, pièce*) et que les mots les plus monosémiques se situent à droite (*Fig, mm, type*), bien qu'ils soient moins bien regroupés. L'axe vertical de la figure 6.4 est donc l'axe de la stabilité, où des positions plus périphériques (en haut et en bas) signifient que les rangs de

monosémie sont moins stables à travers les différentes tailles de fenêtre d'observation. Il est à noter que les mots les plus homogènes et les plus hétérogènes sémantiquement sont les plus stables en ce qui concerne leur rang de monosémie à travers les différentes tailles de fenêtre d'observation.

Afin de mieux comprendre pourquoi les mots en haut (*copeau* et *commande*) et en bas (*axe* et *broche*) se distinguent des autres par leur dissimilarité, il est intéressant d'observer les détails de leur rang de monosémie à travers les 11 tailles de fenêtre d'observation (Cf. figure 6.5).

Ce qu'on observe est que *copeau* et *commande* deviennent plus hétérogènes sémantiquement dans des fenêtres d'observation plus larges. Par contre, *broche* et *axe* deviennent plus homogènes au fur et à mesure que la taille de la fenêtre augmente : leur rang de monosémie s'approche de 1 (sauf dans la fenêtre 3-15). En plus, *Fig* se distancie de *mm* et de *type* dans la visualisation précédente de MDS (Cf. figure 6.4), en raison de son rang beaucoup plus polysémique dans la fenêtre 3-15.

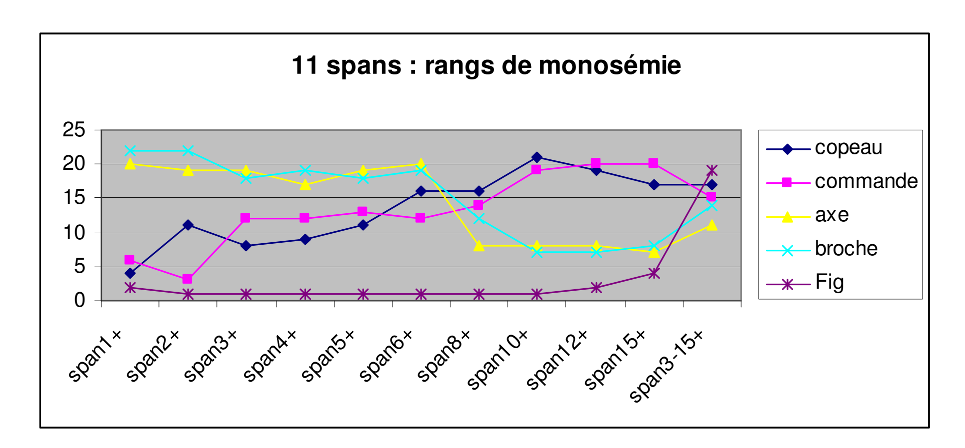


Figure 6.5 Rangs de monosémie dans les 11 fenêtres d'observation

La visualisation ci-dessous (Cf. figure 6.6) est la visualisation la plus intéressante et montre que la taille préconisée de 5 mots à gauche et 5 mots à droite se situe bien au centre des différentes configurations de taille et qu'elle n'est pas périphérique. Les tailles les plus limitées (1 et 2) se trouvent plus à gauche de la visualisation, les tailles les plus importantes (10,12,15) plus à droite. Il est à remarquer également que la taille plus particulière de 3-15 s'avère très périphérique par rapport aux autres.

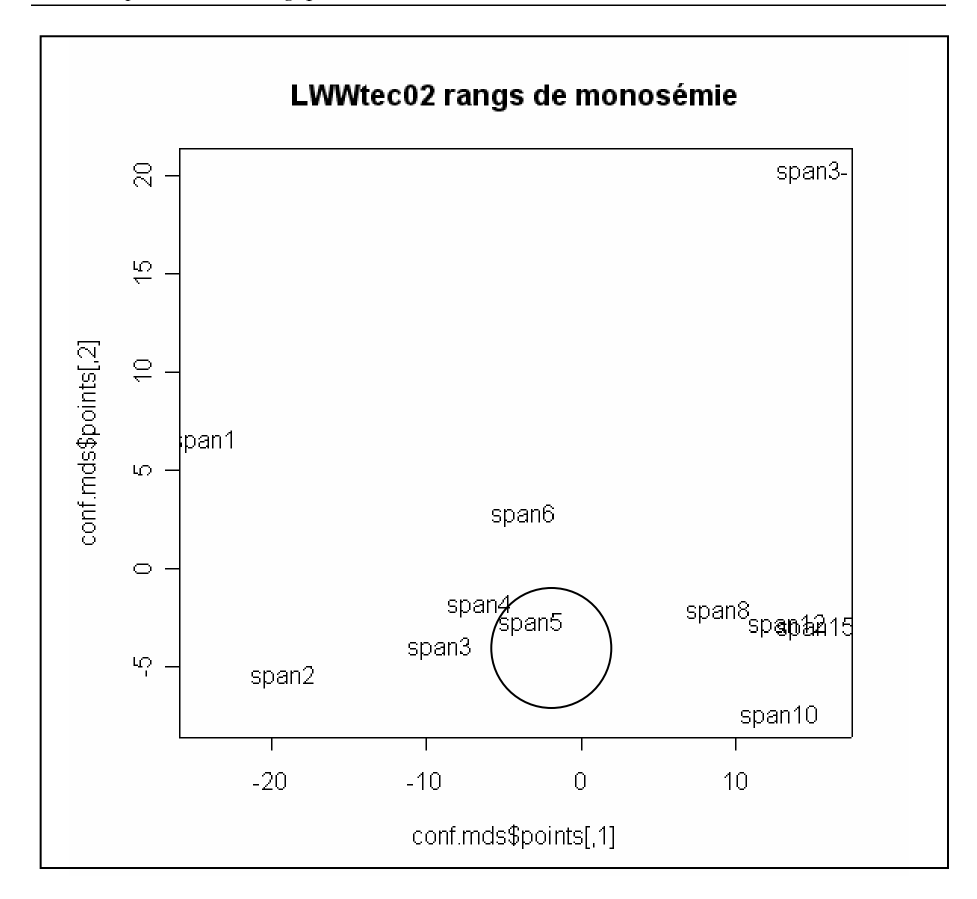


Figure 6.6 Résultat MDS des 11 tailles différentes

Finalement, nous nous proposons d'affiner les résultats du MDS pour les 11 tailles différentes en étudiant le rang de monosémie moyen à travers les 11 configurations de taille, pour chacun des 25 mots. Le rang de monosémie moyen permettra de comparer l'écart-type[145] des 25 mots, ainsi que des 11 tailles différentes.

---

[145] L'écart-type (σ) est la racine carrée de la variance. La variance (σ²) d'une variable est une mesure permettant de voir si les valeurs de la variable sont consistantes entre elles ou si elles varient beaucoup. On calcule la variance en estimant combien, en moyenne, les valeurs de la variable sont déviantes par rapport à la valeur attendue de la variable (la moyenne μ). La variance est donc la moyenne des carrés des écarts à la moyenne μ. L'écart-type σ indique la déviation moyenne de toutes les valeurs par rapport à la moyenne μ.

Le tableau ci-dessous (Cf. tableau 6.6) montre que les mots avec l'écart-type le plus limité sont les mots les plus spécifiques, à savoir *outil*, *machine*, *pièce*. Leur rang de monosémie présente le moins de variation dans les différentes fenêtres d'observation. Par contre, les mots dont le rang de monosémie présente le plus de variation sont *broche*, *axe*, *commande*, *Fig* et *copeau* : leur rang de monosémie varie beaucoup dans les 11 fenêtres d'observation, ce qui confirme les résultats indicatifs visualisés ci-dessus (Cf. figures 6.4 et 6.5).

N°	mot	écart-type
2	*outil*	0,934198733
1	*machine*	1,286291357
4	*pièce*	1,401298099
16	*fraisage*	1,634847783
24	*gamme*	1,921173884
…	…	…
11	*copeau*	5,260487361
9	*Fig*	5,356389557
14	*commande*	5,386852682
10	*axe*	5,671299354
8	*broche*	5,787133063

Tableau 6.6 Ecart-type minimal et maximal des 25 spécificités (pour les 11 tailles)

taille	écart-type
span1	4,966865
span2	4,274313
span3	2,893367
span4	2,545714
span5	**2,371177**
span6	2,624487
span8	2,758248
span10	3,488257
span12	3,152016
span15	3,420889
span3-15	5,27987

Tableau 6.7 Ecart-type des 11 tailles (pour les 25 spécificités)

Pour les 11 tailles de fenêtre d'observation différentes, l'écart-type par taille est également calculé à partir du rang de monosémie moyen par mot[146]. Le tableau ci-dessus (Cf. tableau 6.7) visualise l'écart-type des tailles de fenêtre et confirme la position centrale de celle qui est de taille 5 et qui se caractérise par l'écart-type le plus bas (2,37). Cette fenêtre présente le moins de variations du rang de monosémie des 25 mots par rapport au rang de monosémie moyen pour les 11 tailles. Donc, autrement dit, plus la taille de la fenêtre d'observation est petite ou plus elle est grande, plus le rang de monosémie de chacun de ces 25 mots s'éloigne du rang de monosémie moyen.

6.1.3 Le seuil de significativité

A l'instar des expérimentations précédentes et des analyses de MDS concernant la forme graphique des c et cc et la taille de la fenêtre d'observation, nous procéderons aussi à la comparaison de plusieurs seuils de significativité. Ces expérimentations nous permettront de vérifier si le seuil préconisé (0,9999) n'est pas trop périphérique. Les expérimentations seront de nouveau conduites pour les 25 mots les plus spécifiques, sur le même échantillon, pour la configuration LWWtec02 et pour une fenêtre d'observation de [-5;+5]. Il s'agira des seuils de significativité suivants : 0,95 ($p < 0{,}05$), 0,99 ($p < 0{,}01$), 0,999 ($p < 0{,}001$) et 0,9999 ($p < 0{,}0001$).

Notons que moins on est sévère (seuil de 0,95), plus de c et de cc seront significatifs, et, dès lors, inclus dans la base de données[147]. Par conséquent, plus de cc seront pris en considération pour le calcul du recoupement. Ces cc supplémentaires (car moins significatifs et moins pertinents sémantiquement), pourront soit augmenter le degré de recoupement moyen, s'ils sont identiques à d'autres cc plus significatifs, soit diminuer le degré de recoupement, s'ils sont formellement différents des autres cc plus significatifs.

---

[146] Pour chaque taille, on calcule, pour chaque mot, la différence entre le rang de monosémie du mot pour cette taille (p.ex. la taille 5) et le rang de monosémie moyen du mot, à travers les 11 tailles différentes. Cette différence donne une indication de la déviation du rang pour la taille 5 par rapport au rang moyen. Ces différences pour les 25 mots seront élevées au carré et totalisées pour la taille 5 (et ainsi de suite pour les 11 tailles). Ensuite, la somme pour la taille 5 est divisée par 25 pour connaître la variance. Finalement, la racine carrée de la variance indique l'écart-type des rangs de monosémie dans cette taille de fenêtre d'observation (5).

[147] Par exemple pour le mot *tour* dans la configuration LWWtec02, le nombre de c varie entre 36 (0,9999) et 329 (0,95) et le nombre de cc varie même entre 632 (0,9999) et 21724 (0,95). Pour les détails des différences entre les seuils de 0,9999 et de 0,999 : Cf. annexe 9.2 (Comparaison des seuils de significativité des c à 0.9999 et à 0.999).

La visualisation ci-dessous (Cf. figure 6.7) montre que les seuils de significativité 0,99 et 0,95 (à droite) génèrent des résultats similaires en matière de rang de monosémie pour les 25 spécificités analysées. Ces deux seuils incluent plus de c et de cc et ils se situent loin des deux autres seuils qui incluent des c et des cc plus saillants et plus pertinents. Force est de constater que les deux seuils plus « sévères » (0,999 et 0,9999) se situent à une distance considérable l'un de l'autre, ce qui indique des dissimilarités importantes quant au rang de monosémie.

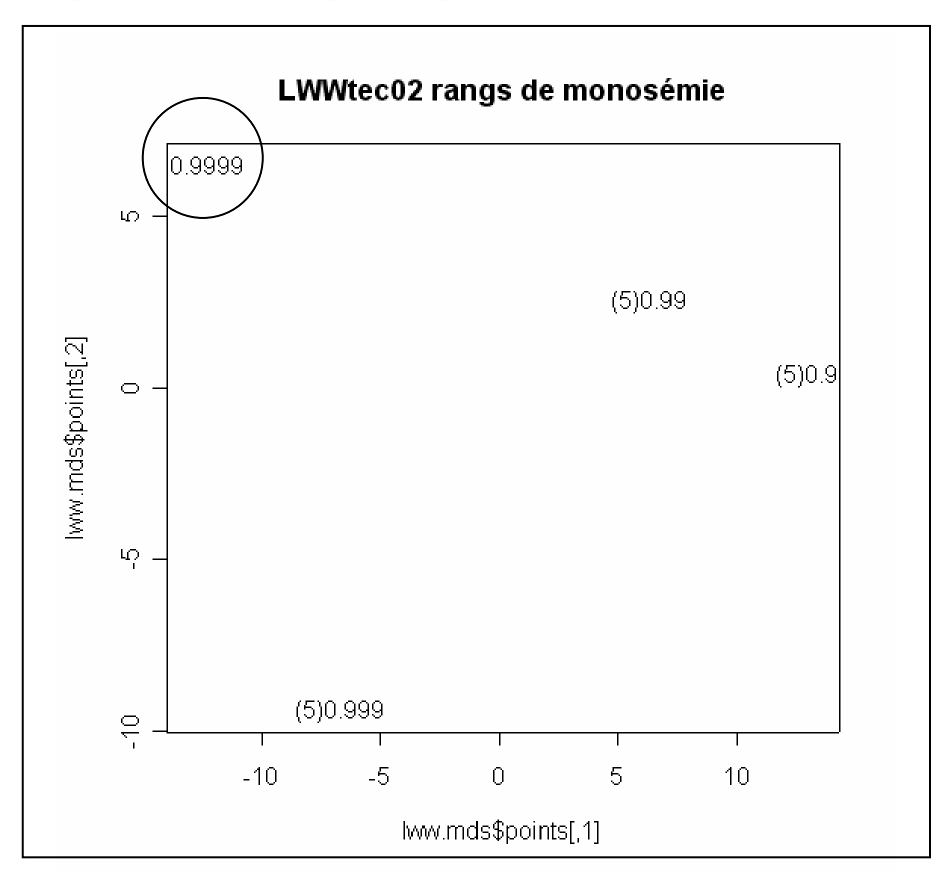


Figure 6.7 Résultat MDS des seuils de significativité

Il est clair que les analyses de MDS devraient faire varier, non seulement le seuil de significativité *ou* la taille de la fenêtre d'observation, mais également plusieurs paramètres à la fois, en l'occurrence le seuil de significativité *et* la taille de la fenêtre d'observation, ou même les trois paramètres analysés ci-dessus.

6.1.4 Analyses faisant varier plusieurs paramètres de configuration

6.1.4.1 La taille de la fenêtre d'observation et le seuil de significativité

Etant donné que nous préférons considérer les cooccurrents et les cooccurrents des cooccurrents au niveau des formes fléchies, parce que sémantiquement plus riches, nous proposons d'inclure dans les analyses de MDS d'abord les deux autres paramètres, à savoir la taille de la fenêtre d'observation et le seuil de significativité. Cette analyse de MDS prendra en considération les 11 tailles (*spans*) (Cf. 6.1.2), à savoir 1, 2, 3, 4, 5, 6, 8, 10, 12, 15 et finalement 3-15, pour les deux seuils de significativité les plus sévères, à savoir 0,9999 et 0,999. Dans le but de ne pas trop compliquer la visualisation de MDS, les tailles au seuil de significativité 0,9999 seront dénommées span+ et celles dont le seuil de significativité est de 0,999 span-, par exemple span1+ au seuil de 0,9999 et span1- au seuil de 0,999.

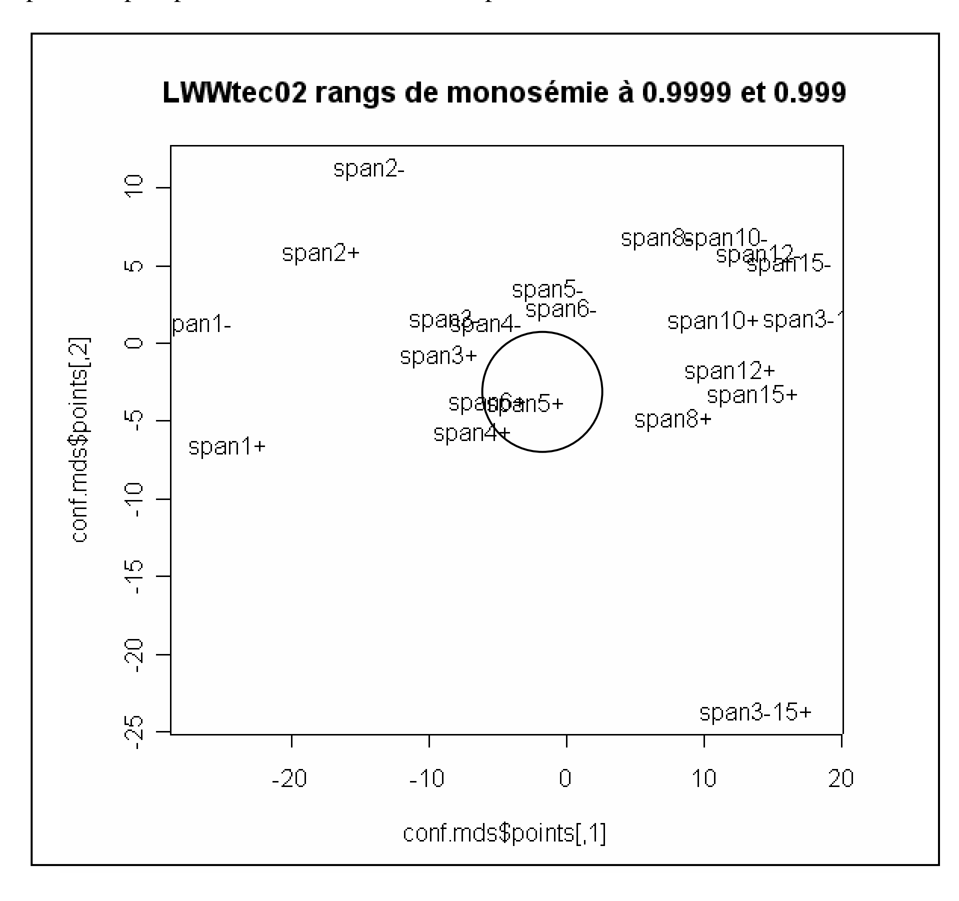


Figure 6.8 Résultat MDS des 2 seuils de significativité et des 11 tailles

La visualisation ci-dessus (Cf. figure 6.8) montre les résultats des analyses qui font varier le seuil et la taille. Signalons que la taille 3-15 au seuil 0,9999 (span3-15+) est très périphérique par rapport aux autres configurations. Il est clair que la taille préconisée [-5;+5] est centrale par rapport aux autres tailles, tant au seuil de 0,999 qu'au seuil plus sévère de 0,9999. Les choix méthodologiques se voient donc confirmés par les résultats de ces analyses MDS qui font intervenir deux paramètres.

6.1.4.2 Trois paramètres de configuration

Les dernières analyses de MDS font varier tous les paramètres de configuration, à savoir la taille de la fenêtre (*span*), le seuil de significativité et la forme graphique ou canonique des c et des cc. Au total, 60 configurations différentes seront envisagées, pour le même échantillon (Ntec02). Les 3 configurations des formes graphiques de c et cc (LWW – LLW – LLL) et les 5 tailles de fenêtre considérées (1-2-3-4-5) fournissent 15 bases de données différentes. Les 4 seuils de significativité pris en considération (0,95-0,99-0,999-0,9999) aboutiront quant à eux à 60 configurations différentes (Cf. tableau 6.8). La numérotation des configurations permet d'afficher de façon plus claire les distances et les (dis)similarités dans l'analyse de MDS. Les configurations les plus intéressantes (span5_0,9999) correspondent aux numéros 17, 37 et 57 ; elles sont marquées en gris clair.

N°	forme	span	seuil	N°	forme	span	seuil	N°	forme	span	seuil
1	LWW	1	0,9999	21	LLW	1	0,9999	41	LLL	1	0,9999
2	LWW	1	0,999	22	LLW	1	0,999	42	LLL	1	0,999
3	LWW	1	0,99	23	LLW	1	0,99	43	LLL	1	0,99
4	LWW	1	0,95	24	LLW	1	0,95	44	LLL	1	0,95
5	LWW	2	0,9999	25	LLW	2	0,9999	45	LLL	2	0,9999
6	LWW	2	0,999	26	LLW	2	0,999	46	LLL	2	0,999
7	LWW	2	0,99	27	LLW	2	0,99	47	LLL	2	0,99
8	LWW	2	0,95	28	LLW	2	0,95	48	LLL	2	0,95
9	LWW	3	0,9999	29	LLW	3	0,9999	49	LLL	3	0,9999
10	LWW	3	0,999	30	LLW	3	0,999	50	LLL	3	0,999
11	LWW	3	0,99	31	LLW	3	0,99	51	LLL	3	0,99
12	LWW	3	0,95	32	LLW	3	0,95	52	LLL	3	0,95
13	LWW	4	0,9999	33	LLW	4	0,9999	53	LLL	4	0,9999
14	LWW	4	0,999	34	LLW	4	0,999	54	LLL	4	0,999
15	LWW	4	0,99	35	LLW	4	0,99	55	LLL	4	0,99
16	LWW	4	0,95	36	LLW	4	0,95	56	LLL	4	0,95
17	LWW	5	0,9999	37	LLW	5	0,9999	57	LLL	5	0,9999
18	LWW	5	0,999	38	LLW	5	0,999	58	LLL	5	0,999
19	LWW	5	0,99	39	LLW	5	0,99	59	LLL	5	0,99
20	LWW	5	0,95	40	LLW	5	0,95	60	LLL	5	0,95

Tableau 6.8 Comparaison des 60 configurations

La figure 6.9 ci-dessous visualise les 20 configurations pour LWWtec02. Elle montre clairement que la configuration préconisée (LWWtec02_span5_0,9999) se situe plutôt en haut de la visualisation (Cf. figure 6.9). Un premier axe d'interprétation serait la diagonale montant du coin gauche inférieur vers le coin droit supérieur. En bas à gauche, sont regroupées les configurations de taille limitée. Au fur et à mesure qu'on monte vers le coin droit supérieur, on retrouve les tailles plus grandes. Un deuxième axe d'interprétation perpendiculaire descend du coin gauche supérieur vers le coin droit inférieur. On observe les seuils les plus sévères pour la taille 3-4-5 en haut de la visualisation et la configuration préconisée au milieu de l'axe horizontal. Elle s'approche le plus de la taille 4 au seuil de 0,999. Le nombre plus élevé de cc se voit ainsi compensé par la taille réduite de la fenêtre d'observation (4). Le fait que la configuration préconisée se trouve en haut de la visualisation (par rapport à l'axe vertical), s'explique par la taille des 20 configurations prises en considération dans cette analyse (de 1 à 5 mais pas au-delà de 5).

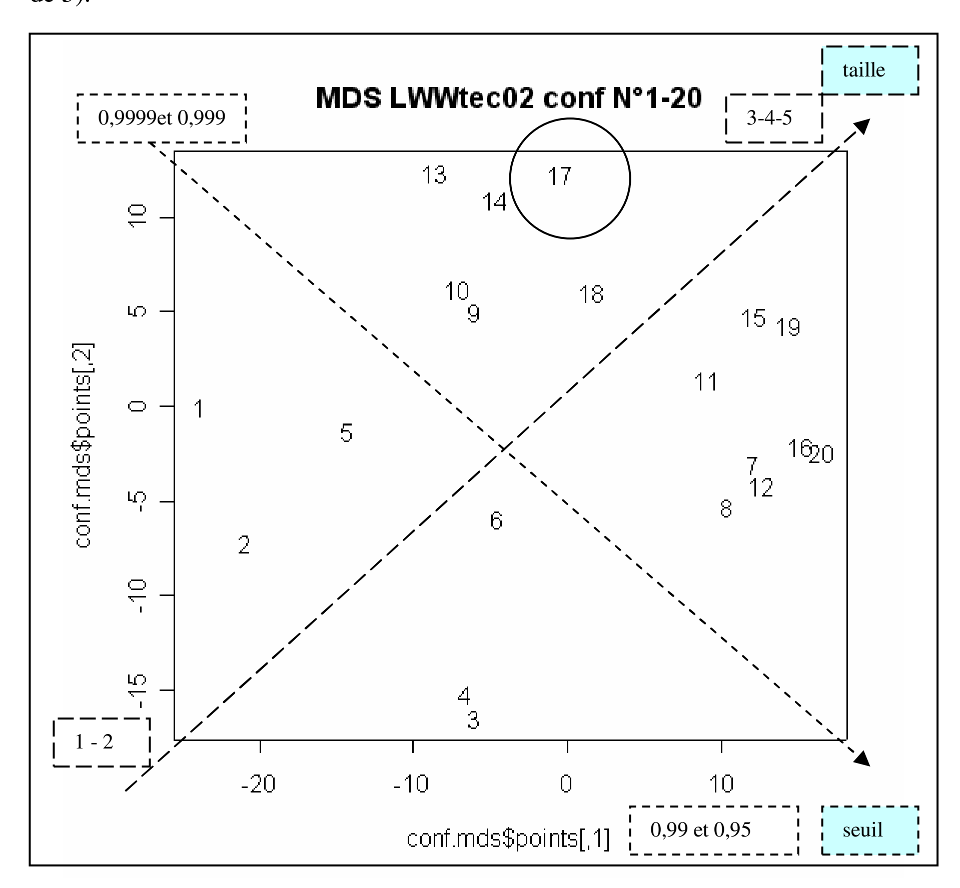


Figure 6.9 Résultat MDS des 20 configurations LWWtec02 (5 tailles et 4 seuils)

Finalement, la figure 6.10 ci-dessous visualise les 60 configurations et leur (dis)similarités[148]. Les configurations des lemmes uniquement (LLL), les numéros 41 à 60, se trouvent dans la moitié inférieure de la représentation, avec les seuils sévères plutôt à gauche. Les mêmes axes d'interprétation diagonaux se dégagent pour les 60 configurations, à quelques exceptions près (53 et 57). La configuration préconisée 17 se caractérise par des similarités avec 13, 14, 9, 10 et 34, c'est-à-dire les tailles 4 et 3 aux seuils 0,9999 et 0,999 (LWW) et la taille 4 au seuil 0,999 (LLW). Elle se situe de nouveau au milieu de la visualisation, en haut.

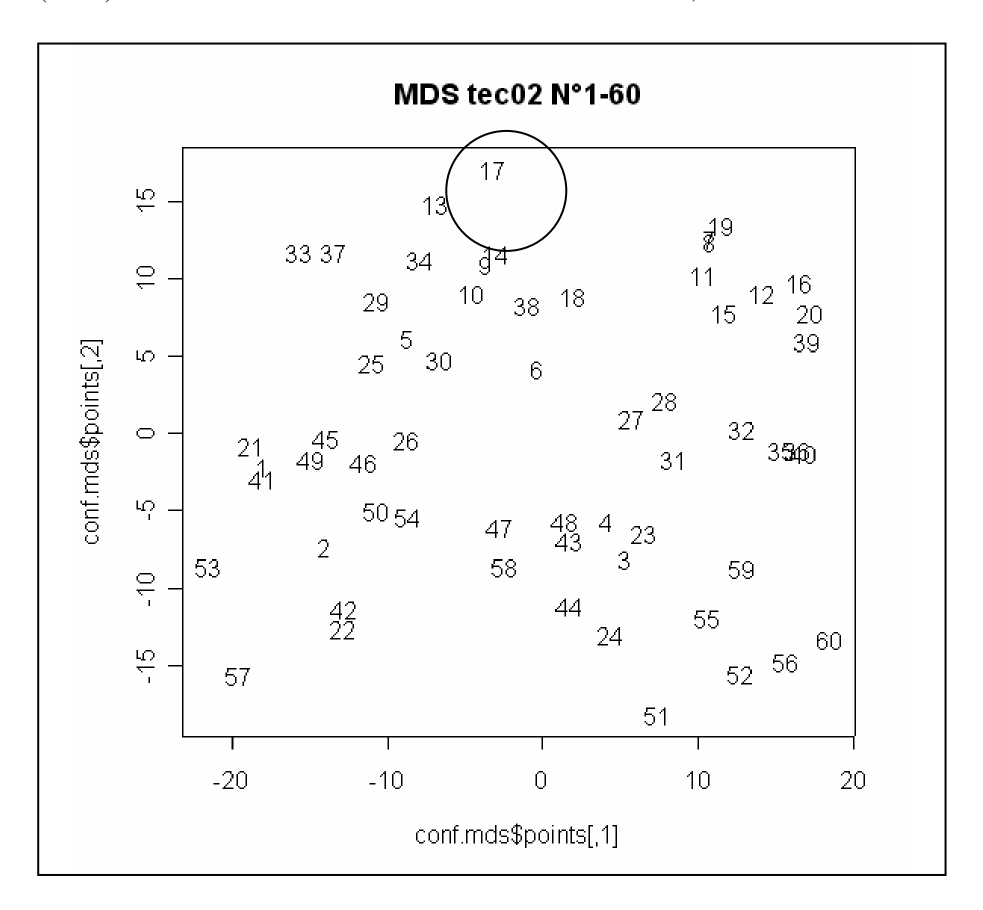


Figure 6.10 Résultat MDS des 60 configurations tec02

[148] Pour des raisons pratiques, nous ne procédons pas à la réalisation de 30 bases de données supplémentaires, pour les tailles (*spans*) 6-8-10-12-15 aux seuils 0,99 et 0,95. Ces seuils n'étant pas très sévères, ils incluent beaucoup de c et de cc (pas toujours très pertinents sémantiquement), ce qui mènerait à des bases de données très lourdes.

En guise de conclusion de toutes les expérimentations et analyses de MDS exploratoires pour les 25 spécificités les plus spécifiques, nous pouvons affirmer que la configuration préconisée (LWWtec02_span5_0,9999) occupe une position bien centrale parmi les différentes configurations. Ces résultats, en termes de rang de monosémie, ressemblent le plus aux résultats de rang de monosémie pour la fenêtre d'observation 4 et aux seuils 0,9999 et 0,999. Même si la configuration préconisée s'avère un peu périphérique sur la visualisation ci-dessus, il est vrai que les visualisations précédentes qui font intervenir des tailles plus importantes, semblent tout de même réserver une position plus centrale à la configuration préconisée.

6.2 FACTEURS DE LA MESURE DE RECOUPEMENT

Les analyses de MDS de la partie précédente montrent que la configuration LWW, dans une fenêtre d'observation de 5 mots à gauche et à droite et à un seuil de significativité de 0,9999, est la configuration la plus stable des différentes expérimentations dans un échantillon. Dans cette partie, les expérimentations seront conduites sur le corpus technique entier, à savoir LWWtecA-F, pour la même configuration (fenêtre d'observation de 5 et seuil de 0,9999). Ces expérimentations visent principalement à expliquer l'importance des facteurs intégrés dans la mesure de recoupement et à les caractériser, dans le but de mieux comprendre le fonctionnement et la sensibilité de la mesure de recoupement. En effet, plusieurs questions se posent. Faut-il tenir compte du nombre de c, du nombre de cc par c ou des cc isolés ? La première section de cette partie sera consacrée à l'importance du nombre de cooccurrents (ou c) (6.2.1). Les sections suivantes analyseront les cooccurrents des cooccurrents, leur recoupement d'une part (6.2.2) et leur fréquence d'autre part (6.2.3). Finalement, la dernière section discutera la sensibilité de la mesure de recoupement (6.2.4).

6.2.1 L'importance du nombre de cooccurrents (c)

Comme nous l'avons expliqué dans le chapitre précédent, la mesure de recoupement permettra de quantifier la monosémie en l'implémentant comme homogénéité sémantique. Rappelons que la mesure repose sur le nombre de cooccurrents, le nombre de cooccurrents des cooccurrents et la fréquence des cooccurrents des cooccurrents (Cf. figure 6.11).

$$\sum_{cc} \frac{\text{fq}\, cc}{\text{nbr total}\, c \cdot \text{nbr total}\, cc}$$

Figure 6.11 Mesure de recoupement (Cf. figure 5.2)

La mesure de recoupement détermine à quel point les cc se recoupent ou à quel point ils sont partagés par les c. On peut se demander s'il est important de tenir compte, dans le dénominateur de la formule, du nombre de c et quel serait l'impact sur le rang de monosémie, si ce facteur était exclu de la formule.

6.2.1.1 Comparaison de plusieurs mesures

Afin de vérifier l'impact des cooccurrents (c) sur le rang de monosémie des spécificités, nous proposons de comparer plusieurs mesures. Les expérimentations seront conduites sur le corpus technique entier, pour un petit échantillon de 50 spécificités, représentatives des 4717 spécificités de l'ensemble. L'échantillon comprend des mots intuitivement polysémiques, tels que *machine* et *tour*, des mots intuitivement monosémiques, tels que *Fig, m* et *m/min*, des mots très fréquents, moins fréquents et très peu fréquents dans le corpus technique, ainsi que des mots très spécifiques, moins spécifiques et très peu spécifiques du corpus technique (Cf. tableau 6.9). Cette diversité d'homogénéité, de fréquence et de spécificité permettra de vérifier à fond l'impact des différentes mesures sur le degré et le rang de monosémie de l'échantillon.

spécificité	fq1	LLR
machine	12671	50521,91
outil	8306	32037,72
Fig	2680	12194,00
arête	1870	8213,91
précision	2263	7663,01
usiner	1577	7045,52
système	4052	6915,85
permettre	4883	5848,03
m	1240	5641,04
avance	1832	5200,57
effectuer	1508	2428,40
puissance	1354	2409,23
Iso	516	2347,19
abrasif	523	2332,91
table	1141	2248,02
électroérosion	481	2173,86
travail	3211	1879,94
etc	927	1334,09
m/min	293	1332,75
technique	1284	1283,85
emboutissage	288	1268,64
tour	1476	748,78
meulage	115	523,07
variable	268	522,68
valeur	987	522,24
concept	368	521,38
tonne	398	519,60
fraiser	114	518,52
commander	249	191,06
externe	131	191,02
raboutage	42	191,01
mm/s	42	191,01
assembler	103	189,82
numériquement	17	52,23
réfrigération	17	52,23
verre	116	52,18
présérie	13	52,14
endommagement	13	52,14
maîtriser	120	52,13
collaboration	104	21,65
cloison	18	21,55
insérer	40	21,36
extérieurement	9	20,53
numérotation	9	20,53
réutilisable	9	20,53
microbiologique	9	20,53
puisard	3	9,36
commuter	3	9,36
vidangeur	3	9,36
batch	3	9,36

Tableau 6.9 Echantillon de 50 spécificités représentatives

Plusieurs mesures seront comparées, généralement des variations sur le thème de la mesure de recoupement de base ou la mesure de monosémie (Cf. figure 6.11).

(1) M_monosémie : ∑ (fq cc / (nbr total c * nbr total cc))[149] (Cf. figure 6.11)

(2) M_cc_diff : -log (nbr cc différents / nbr total cc)[150]

(3) M_fq_cc : ∑ (fq cc / nbr total cc)[151]

(4) M_c/2 : ∑ (fq cc / ((nbr total c / 2) * nbr total cc))[152]

(5) M_cc-types : ∑ (fq cc-t / (nbr total c * nbr total cc-t))[153]

= nbr total *cc* / (nbr total *c* * nbr *cc* différents)

Du point de vue méthodologique, la mesure de monosémie, M_monosémie (1), s'oppose aux deux mesures suivantes, à savoir M_cc_diff (2) et M_fq_cc (3), parce que ces deux mesures ne tiennent pas compte du nombre total de c.

La quatrième mesure, M_c/2 (4), se caractérise par le fait que le nombre de c est divisé par deux, ce qui permet également d'évaluer l'impact du nombre de c. Or, cet impact sera plus limité en raison de la pondération (division par deux). Finalement, la dernière mesure, M_cc-types (5), se situe au niveau des cc-types (cc différents ou uniques), étant donné qu'elle tient compte de la fréquence des cc-types (*cc-t*) et du nombre total de cc-types (*cc-t*), au lieu des occurrences de cc (*cc-tokens*) de la

[149] Il est clair qu'on somme sur tous les cc (*cc-tokens*) (Cf. formule détaillée : figure 6.11).

[150] Le nombre de cc différents (*cc-types*) est divisé par le nombre total de cc (*cc-tokens*). Le résultat de cette fraction se situe toujours entre 0 et 1. Plus il est près de 1, plus il y a de cc différents et moins les cc se recoupent. Plus il est près de 0, plus les cc se recoupent. Ensuite, *-log* (fraction) permet d'aboutir à des valeurs tendant vers l'infini pour les plus monosémiques (fraction près de 0) et à des valeurs s'approchant de zéro pour les plus polysémiques (fraction près de 1). Ainsi, les degrés de monosémie de cette mesure pourront être classés par ordre décroissant pour obtenir les rangs de monosémie.

[151] Cette mesure ressemble beaucoup à la mesure de recoupement de base, mais elle n'inclut pas le nombre total de c dans le dénominateur. On somme sur tous les cc (*cc-tokens*).

[152] On somme sur tous les cc (*cc-tokens*).

[153] On somme sur les cc différents (*cc-types ou cc-t*). Il est à noter que la somme de la fréquence des cc différents (*cc-types*) égale le nombre total de cc (*cc-tokens*).

mesure de base M_monosémie (1). La somme de la fréquence de tous les cc-types équivalant au nombre total de cc (occurrences ou *cc-tokens*), la mesure équivaut à la formule simplifiée mentionnée plus bas.

Pour les 50 spécificités[154] de l'échantillon, les cinq mesures alternatives permettent de générer cinq rangs de monosémie par spécificité ; le rang de monosémie de la mesure de monosémie de base (1) est indiqué en gris clair (Cf. tableau 6.10).

N°	spécificité	(1)	(2)	(3)	(4)	(5)
1	*machine*	49	1	2	49	49
2	*outil*	48	2	6	48	48
3	*permettre*	44	5	12	44	43
4	*système*	46	8	8	46	47
5	*travail*	43	13	16	43	40
6	*Fig*	30	3	1	30	46
7	*précision*	29	14	11	29	32
8	*arête*	37	17	14	37	37
9	*avance*	32	4	9	32	36
10	*usiner*	36	19	20	36	30
11	*effectuer*	33	10	19	33	28
12	*tour*	45	20	13	45	45
13	*puissance*	39	9	10	39	38
14	*technique*	47	18	18	47	41
15	*m*	38	7	3	38	44
16	*table*	41	12	15	41	39
17	*valeur*	40	15	17	40	33
18	*etc*	35	33	21	35	31
19	*abrasif*	42	21	22	42	35
20	*Iso*	31	16	4	31	42
21	*électroérosion*	27	22	24	27	25
22	*tonne*	22	11	5	22	34
23	*concept*	21	25	27	21	20
24	*m/min*	19	6	7	19	29
25	*emboutissage*	26	23	23	26	27
26	*variable*	24	30	25	24	23
27	*commander*	23	27	33	23	21
28	*externe*	28	32	29	28	24

[154] Il est à noter que le mot *endommagement* (n° 42) n'entraîne pas de résultats pour le calcul du degré de recoupement, parce que cette spécificité n'a pas de c au seuil de significativité 0,9999.

29	*maîtriser*	14	39	40	14	13
30	*verre*	25	37	35	25	22
31	*meulage*	34	31	26	34	26
32	*fraiser*	17	44	44	17	16
33	*collaboration*	18	36	31	18	19
34	*assembler*	20	28	34	20	18
35	*raboutage*	15	38	36	15	15
36	*mm/s*	16	26	28	16	17
37	*insérer*	11	34	37	11	9
38	*cloison*	12	35	32	12	14
39	*numériquement*	10	41	39	10	10
40	*réfrigération*	5	47	46	5	6
41	*présérie*	13	42	41	13	11
42	*endommagement*	--	--	--	--	--
43	*extérieurement*	1	24	38	1	1
44	*numérotation*	4	45	45	4	4
45	*réutilisable*	6	46	47	6	5
46	*microbiologique*	9	29	30	9	12
47	*puisard*	8	40	42	8	8
48	*commuter*	7	49	49	7	7
49	*vidangeur*	3	43	43	3	3
50	*batch*	2	48	48	2	2

Tableau 6.10 Echantillon de 50 spécificités : rangs alternatifs de monosémie

Premièrement, on observe que la mesure de monosémie de base, M_monosémie (1), marquée en gris clair dans la troisième colonne, accorde des rangs de monosémie entre 1 et 10 à des mots peu fréquents et peu spécifiques, visualisés en bas de liste. Pour des raisons évidentes, les mots les moins fréquents auront moins de chances d'apparaître dans des contextes sémantiquement très hétérogènes. Comme les deux mesures M_cc_diff (2) et M_fq_cc (3) ne prennent pas en compte le nombre de c, elles accordent les rangs de monosémie les plus bas (grosso modo entre 1 et 15) aux mots les plus fréquents, qui correspondent toutefois à des mots intuitivement plutôt hétérogènes sémantiquement, tels que *machine* et *outil*. Pour les mots intuitivement plutôt monosémiques, par contre, les rangs de monosémie accordés par ces deux mesures alternatives correspondent bien à l'intuition (3 et 1 pour *Fig*, 7 et 3 pour *m* et 6 et 7 pour *m/min*).

Deuxièmement, les rangs de monosémie de la mesure M_c/2 (4) correspondent parfaitement aux rangs de monosémie de la mesure de monosémie de base (1). Par conséquent, le fait d'inclure un facteur de pondération pour le nombre de c n'affecte en rien les rangs de monosémie. Les degrés de monosémie, par contre, sont bel et bien modifiés, car ils sont plus élevés pour toutes les spécificités, étant donné que le

dénominateur de la mesure est un nombre moins élevé. Toutefois, l'intégration du facteur de pondération dans la mesure (4) entraîne une conséquence méthodologique et mathématique très importante, parce que le résultat ne se situe plus entre 0 et 1. Dès lors, il sera difficilement interprétable (Cf. chapitre 5). Le résultat pourra aussi dépasser 1, ce qui est d'ailleurs le cas pour les mots les plus monosémiques, tels que *extérieurement*. Pour cette raison, nous n'adoptons pas la mesure alternative M_c/2 (4) pour les analyses sémantiques définitives. L'expérimentation pour cette mesure visait uniquement à vérifier l'impact des facteurs pondérés sur les rangs de monosémie dans l'échantillon de 50 spécificités et à mieux comprendre les facteurs inclus dans la formule de base (1).

Troisièmement, la mesure M_cc-types (5) repose sur les *cc-types* et non sur les *cc-tokens*. Méthodologiquement, le recoupement d'un cc-type (fréquence de ce cc-type) pèse moins lourd sur le résultat final que le recoupement d'un *cc-token*, car il est compté une fois, alors que le recoupement du cc-token sera compté autant de fois que la fréquence du cc-token[155]. La comparaison des résultats, c'est-à-dire des rangs de monosémie des deux mesures (1) et (5) montre peu de différences, à première vue. En effet, les mots hétérogènes sémantiquement dans (1) le sont également dans (5), ce qui est visualisé dans la dernière colonne. Toutefois, les mots intuitivement monosémiques, tels que *Fig, Iso* et *m/min*, se retrouvent à des rangs considérablement plus polysémiques (car plus élevés et plus près de 50) pour la mesure M_cc-types (5) : *Fig* (30 vs. 46), *Iso* (31 vs 42), *m/min* (19 vs 29). Intuitivement, les résultats de la mesure (5), en termes de rangs de monosémie, sont donc moins plausibles que les résultats de la mesure de base (1), ce qui s'explique par la façon de calculer le recoupement des cc (*cc-tokens* (1) vs. *cc-types* (5)).

Reprenons finalement les deux mesures M_cc_diff (2) et M_fq_cc (3) qui ne prennent pas en considération le nombre de c, afin d'expliquer pourquoi ils produisent des résultats contre-intuitifs. La mesure (2) est basée sur le rapport entre le nombre de cc uniques (ou différents) (=*cc-types*) et le nombre total de cc (=*cc-tokens*). Ce rapport augmente, si les mots sont moins spécifiques et moins fréquents et s'ils ont moins de c et de cc. Même si le nombre total de cc d'un mot de base augmente (*cc-tokens*), le nombre de cc différents (*cc-types*) de ce mot n'augmentera pas dans la même mesure, ce qui correspond grosso modo au TTR (*Type-Token Ratio*) du vocabulaire d'un corpus (Cf. chapitre 3). En effet, si le nombre

---

155 Dans la mesure de base (1), un cc qui figure trois fois dans la liste des cc du mot de base (fréquence 3) sera comptabilisé 3 fois à cette fréquence 3, ce qui donne lieu à un facteur de 3² (= 9) dans le numérateur de la formule. Par contre, dans la mesure alternative des cc-types (5), un cc qui figure trois fois (fréquence 3) sera comptabilisé 1 fois à la fréquence 3 (seulement le cc unique ou cc-type), ce qui donne lieu à un facteur de 3 dans le numérateur de la formule.

d'occurrences (*tokens*) augmente, le nombre de types (*types*) n'augmente pas dans la même mesure. Ainsi, deux cas de figure se distinguent (Cf. tableau 6.11) pour la mesure M_cc_diff (2).

Mots plutôt fréquents	Mots peu fréquents
mots très spécifiques	mots peu spécifiques
mots sémantiquement hétérogènes	mots sémantiquement homogènes
nombre élevé de c et de cc	nombre limité de c et de cc
moins de cc différents par rapport au nombre total de cc	plus de cc différents par rapport au nombre total de cc
machine : 23163 *cc-tokens* et 9027 *cc-types* : 38%	*batch* : 73 *cc-tokens* et 72 *cc-types* : 97%
moins de cc nouveaux si le nombre total de cc augmente	plus de cc nouveaux si le nombre total de cc augmente
plus de cc partagés (au moins 2 fois)	moins de cc partagés (au moins 2 fois)
théoriquement PLUS de chances de recoupement (tendance à la monosémie)	théoriquement MOINS de chances de recoupement (tendance à la polysémie)

Tableau 6.11 Cas de figure : nombre de cc différents et nombre total de cc

Si un cc est partagé par 2 c sur 390 c (par exemple pour *machine*), il est bel est bien partagé et il n'est pas unique, mais pour l'image globale de ce mot, c'est une très faible indication de monosémie. Si, par contre, un cc est partagé par 2 c sur 2 ou 3 c, il est aussi partagé (pas unique), mais pour l'image globale de ce mot, c'est une plus forte indication de monosémie. Autrement dit, la polysémie obtenue en regardant uniquement le nombre de cc uniques ou différents n'est qu'apparente. Par conséquent, il faut également tenir compte du nombre de fois que chaque cc est partagé, donc du nombre de c ou de cooccurrents avec lesquels il apparaît. En effet, il faut inclure le nombre total de c, car l'exclure revient à la mesure (3) et génère également des résultats peu intuitifs.

De ce qui précède, il ressort que la mesure de recoupement de base (1) est une mesure plus intuitive, en dépit du fait que les mots peu fréquents (fréquence absolue de 3 ou de 9 dans le corpus technique), se voient attribuer des rangs de monosémie inférieurs à 10. Il s'ensuit que les mots peu fréquents dans le corpus technique et sémantiquement homogènes relèguent les autres mots intuitivement plutôt monosémiques, tels que *Fig, m, m/min*, à des rangs un peu plus polysémiques (19 ou 30). Force est de constater que la mesure de recoupement de base accorde les rangs les plus polysémiques (entre 40 et 50 dans cet échantillon) aux mots les plus fréquents ayant beaucoup de c. C'est une première indication que la mesure de recoupement semble être sensible à la fréquence absolue de la spécificité dans le corpus technique ainsi qu'à son nombre de c (Cf. 6.2.4).

La mesure de recoupement a fait l'objet d'une validation manuelle à partir de l'analyse manuelle des collocations (Cf. annexe 9 : tableaux A9.1-2-3). Rappelons que des collocations sont des combinaisons fixes et récurrentes d'un mot de base (spécificité) et d'un cooccurrent très pertinent (Cf. chapitre 5). Nous avons relevé tous les cooccurrents statistiquement les plus pertinents (au seuil de significativité le plus sévère de 1). Ce seuil sévère permet donc de limiter le nombre de cooccurrents pour l'analyse manuelle. Les cooccurrents les plus pertinents ont été identifiés pour les mots de base suivants : *machine, outil, tour, avance, arête, m/min, Iso* (Cf. annexe 9). Pour les mots hétérogènes sémantiquement, *machine, outil, tour, avance* et *arête,* l'hétérogénéité des cooccurrents statistiquement très significatifs reflète effectivement celle du mot de base. Ainsi, on retrouve pour *tour*, d'une part *minute, mille* (sens : « rotation, révolution ») et d'autre part *centre, horizontal, bi-broche, …* (sens : « machine-outil pour l'usinage de pièces »). Il est à noter que pour *machine*, les unités polylexicales se manifestent clairement à travers les cooccurrents très significatifs (*machine + à + meuler / scier / rectifier*).

Nous avons également procédé à une validation externe de notre mesure de recoupement au moyen de dictionnaires, puisque nous ne disposons pas de listes de sens préétablis, ni de *Gold Standard*, ni d'autres mesures sémantiques similaires. Les résultats détaillés sont visualisés dans le document en annexe (Cf. annexe 9 : tableau A9.4) : ils confirment les résultats de notre mesure de monosémie pour l'échantillon des 50 spécificités représentatives. Il convient de signaler que les mots les plus fréquents, tels que *machine* et *outil*, entrent très souvent dans la composition d'unités polylexicales (*machine à fraiser, machine à usiner, …*), ce qui pourrait en partie expliquer leur hétérogénéité sémantique[156]. Comme nous l'avons évoqué ci-dessus, les unités polylexicales constituent une piste de recherche très intéressante et complémentaire de notre recherche, qui se limite aux unités simples. Des recherches ultérieures permettront certainement d'approfondir la sémantique des unités polylexicales, mais ces recherches dépassent le cadre méthodologique que nous nous sommes fixé dans notre thèse de doctorat.

---

[156] Les mots spécifiques se retrouvent en grande partie dans la liste des mots les plus fréquents du corpus technique. Ils entrent souvent dans la composition des syntagmes nominaux et des unités polylexicales (terminologiques et monosémiques). Comme ces unités polylexicales ont des distributions hétérogènes, il pourrait en résulter que les mots les plus fréquents (et constituants des unités polylexicales) aient des cooccurrents très différents. La polysémie des mots les plus spécifiques et les plus fréquents pourrait donc s'expliquer par le fait qu'ils entrent dans la composition de nombreuses unités polylexicales. A ce sujet, il serait également intéressant de s'interroger sur la sémantique des termes réduits (Cf. Jacques 2003) et sur les phénomènes de coréférence. Ainsi, un mot simple pourrait constituer la reprise anaphorique d'une unité polylexicale, par exemple *cette machine* qui reprend *machine à usiner*.

6.2.1.2 L'impact du seuil de significativité des cooccurrents

Il est clair que les c ou les cooccurrents jouent un rôle fondamental dans la formule de la mesure de recoupement (Cf. 6.2.1.1). En plus, la partie précédente sur la configuration la plus stable (Cf. 6.1) a démontré l'importance du seuil de significativité des c et des cc ; il se peut par exemple qu'ils soient tous les deux à 0,999 ou tous les deux à 0,9999. Le seuil le plus sévère permet bien entendu de repérer les cooccurrents (c) et les cooccurrents des cooccurrents (cc) les plus pertinents, car un seuil moins sévère risque de générer plus de bruit. Toutefois, la question se pose de savoir quel sera l'impact sur les rangs de monosémie, si le seuil des cooccurrents varie entre 0,9999 et 0,999, mais si le seuil des cc est maintenu à 0,9999. La même question se pose si on maintient les c au seuil de 0,9999, tout en faisant varier le seuil des cc (Cf. annexe 9).

- Configuration de base : c 0,9999 et cc 0,9999
- Configuration intéressante de c : c 0,999 et cc 0,9999
- Configuration informative de cc : c 0,9999 et cc 0,999

Les listes et les résultats des expérimentations à ce sujet (Cf. annexe 9) montrent que le changement de seuil pour les c (c'est-à-dire le changement de c 0,9999 à c 0,999) affecte surtout le degré de recoupement et le rang de monosémie des mots les moins fréquents et les moins spécifiques. Beaucoup de c se rajoutent, si on est moins sévère (c 0,999 et cc 0,9999). Ces mots généralement n'ont que 20 à 30% des c au seuil 0,9999 par rapport au seuil 0,999. Ils apportent également des cc au seuil 0,9999, ce qui entraîne des changements de degré de recoupement et dès lors, des différences de rang de monosémie. Ces dernières sont importantes (différence de rang de 3000 ou de 2000) et elles sont négatives, ce qui veut dire que les mots affectés (peu fréquents et peu spécifiques) deviennent plus hétérogènes sémantiquement si on passe à la configuration intéressante de c (c 0,999 et cc 0,9999) : plus de c se rajoutent, plus on introduit de l'hétérogénéité potentielle. Les mots les plus fréquents et les plus spécifiques se caractérisent par l'hétérogénéité sémantique, tant dans la configuration de base (c 0,9999 et cc 0,9999) que dans la configuration intéressante de c (c 0,999 et cc 0,9999). S'il y a des différences de rang de monosémie, elles sont positives et limitées. Dans la configuration intéressante de c (c 0,999 et cc 0,9999), les mots les plus fréquents et les plus spécifiques sont un peu moins hétérogènes, donc il y a un peu plus de recoupement.

Si l'on contrôle pour le seuil des c (en le maintenant à 0,9999) et si l'on fait varier uniquement le seuil des cc (de 0,9999 à 0,999), le nombre de c reste égal, mais il y a des cc qui se rajoutent, parce qu'on est moins sévère pour les cc (0,999). Dans cette configuration informative de cc (c 0,9999 et cc 0,999), on observe beaucoup moins

de différences de rang de monosémie et, ce qui plus est, des différences moins importantes (différence de rang de 1155 au maximum). Ceci indique clairement que c'est surtout le seuil de significativité des c qui influence le degré de recoupement et dès lors le rang de monosémie. Ces expérimentations tendent donc également à confirmer l'importance du « nombre total de c » dans la formule de la mesure de recoupement.

6.2.2 Le recoupement des cooccurrents des cooccurrents (cc)

La section précédente (6.2.1) a permis de prendre conscience de l'importance, dans la mesure de recoupement, du nombre de c ou cooccurrents du mot de base (spécificité). Dans cette section, nous nous interrogeons sur le nombre de cc par c et sur le recoupement des cc par paire de c.

6.2.2.1 La longueur des vecteurs-cc

Pour des raisons pratiques, nous proposons d'introduire la notion de « vecteur-cc ». Un vecteur-cc regroupe l'ensemble des cc par c, c'est-à-dire l'ensemble des collocatifs (ou cooccurrents) du cooccurrent du mot de base. Chaque c représente donc un vecteur-cc. La longueur d'un vecteur-cc indique le nombre de cc pour ce c, donc le nombre de cc qui sont inclus dans le vecteur-cc. De telle façon, on pourra aussi analyser la longueur de tous les vecteurs-cc, c'est-à-dire le nombre de cc par c, pour tous les c, ainsi que la distribution des longueurs des vecteurs-cc. Il est à noter que chaque cc pourra apparaître une fois par c. En effet, par c ou par vecteur-cc, il s'agit de types de cc (*cc-types*). Toutefois, en regardant les cc de tous les c ou de tous les vecteurs-cc d'un mot de base, le même cc pourra apparaître plusieurs fois ou appartenir à plusieurs vecteurs-cc. Il s'agit donc pour ce mot de base d'occurrences de cc (ou de *cc-tokens*).

Des scripts en Python permettent de définir des fonctions qui génèrent (1) la longueur des vecteurs-cc par spécificité, c'est-à-dire le nombre de cc pour chaque c de cette spécificité et (2) le nombre de vecteurs-cc d'une certaine longueur, ce qui permet d'étudier la distribution des longueurs des vecteurs-cc.

Ainsi, la spécificité *machine*, par exemple, se caractérise par un nombre très important de vecteurs-cc (390) dont 123 longueurs sont différentes (Cf. figure 6.12). La plupart des vecteurs-cc sont d'une longueur de 11, 12 ou 14, ce qui signifie que la plupart des c ont entre 11 et 14 cc (au seuil de significativité de 0,9999). Toutefois, il y a aussi des vecteurs-cc très courts, de longueur 3 ou 4 (ces c ont 3 ou 4 cc différents) ainsi que des vecteurs-cc extrêmement longs, de longueur 739 par exemple ou même de 854, ce qui est assez étonnant. En analysant les longueurs des vecteurs-cc de *machine*, le mot le plus fréquent et le plus spécifique et donc en analysant le nombre de cc pour chaque c, on se rend compte de la distribution

asymétrique des longueurs (Cf. figure 6.12) : il y a énormément de longueurs différentes, surtout pour les longueurs extrêmement longues.

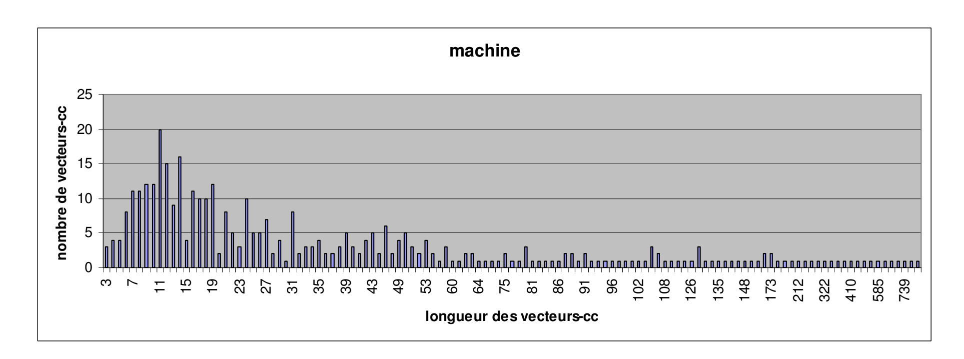


Figure 6.12 Distribution des longueurs des vecteurs-cc (machine)

Bien entendu, la question se pose de savoir quels sont ces vecteurs-cc extrêmement longs et à quels c ils appartiennent. Pour *machine*, les vecteurs-cc les plus longs regroupent les cc des cooccurrents suivants : '*des*' 854, '*les*' 780, '.' 739. Pour la spécificité *tour*, les vecteurs-cc les plus longs caractérisent les cooccurrents suivants : '*à*' 688 et '*un*' 613.

Ces vecteurs-cc extrêmement longs qui correspondent souvent à des mots grammaticaux (articles, déterminants, prépositions, etc.) ne sont pas très pertinents sémantiquement : leurs cc ne sont pas tous porteurs de sens. Par conséquent, nous avons procédé à deux expérimentations qui consistent à limiter le nombre de cc par vecteur-cc à 250. D'une part, pendant le calcul du recoupement, on a tenu compte uniquement des premiers 250 cc par vecteur-cc (aléatoires), ce qui revient à couper la queue des vecteurs-cc extrêmement longs. D'autre part, les vecteurs-cc d'une longueur supérieure à 250 n'ont pas été pris en considération pour le calcul du recoupement, ce qui consiste à exclure les vecteurs-cc extrêmement longs. Le fait de limiter à 250 le nombre de cc de ces vecteurs-cc extrêmement longs ou de les exclure complètement n'affectera sans doute pas l'analyse, car leur apport sémantique est limité. D'ailleurs, un c avec 668 cc significatifs n'indique pas de sens dominant. Ces deux expérimentations visent également à vérifier si la limitation du nombre maximal de cc par c-vecteur permet d'aboutir à une comparaison plus fiable des vecteurs-cc en matière de recoupement. Les vecteurs-cc extrêmement longs risquent en effet de ne pas donner l'occasion de recoupement aux vecteurs-cc plus (ou très) courts et risquent donc de déformer les résultats.

Pour les vecteurs-cc extrêmement longs, deux cas de figure sont possibles théoriquement.

- Soit, il y aurait moins de recoupement pour ces vecteurs-cc, car ils contiennent énormément de cc différents et surtout des cc isolés, qui ne coïncident pas forcément formellement avec les cc des autres vecteurs-cc plus courts.

- Soit, il pourrait y avoir plus de recoupement pour ces vecteurs-cc extrêmement longs, car plus ils ont de cc, plus ils peuvent avoir des cc en commun avec d'autres vecteurs-cc.

La comparaison des résultats en termes de degré de recoupement et de rang de monosémie (Cf. annexe 9) permet de tirer les conclusions suivantes. Généralement, le degré de recoupement des 50 spécificités diminue, tant pour le maximum de 250 cc (*max250*) que pour les vecteurs-cc inférieurs à 250 cc (*under250*). Cette diminution du degré de recoupement indique que les mots deviennent plus hétérogènes sémantiquement et que les cc supprimés ou exclus sont responsables d'un certain recoupement. Toutefois, la diminution générale du degré de recoupement ne se traduit pas pour autant par de grandes différences en matière de rang de monosémie, étant donné que la plupart de ces mots subissent les mêmes tendances pour les deux expérimentations (coupe à 250 cc (*max250*) et exclusion des vecteurs-cc > 250 (*under250*)) (Cf. annexe 9). On observe tout de même que les mots les plus spécifiques et les plus fréquents, avec le plus de c (donc le plus de vecteurs-cc) et le plus de cc, ont le plus souvent des vecteurs-cc d'une longueur supérieure à 250, qui sont soit coupés soit exclus.

Les mêmes expérimentations de coupe (*max250*) et d'exclusion (*under250*) pour toutes les 4717 spécificités permettent d'évaluer les différences en matière de corrélation entre le rang de spécificité et le rang de monosémie[157] (Cf. chapitre 7). Comme le rang de monosémie est très peu affecté par les opérations de coupe ou d'exclusion des vecteurs-cc extrêmement longs, ceux-ci pourront être maintenus sans problèmes méthodologiques. Et donc les analyses sémantiques du calcul de recoupement pourront prendre en considération tous les cc statistiquement significatifs au seuil de significativité choisi de 0,9999, y compris les mots grammaticaux[158].

---

[157] Ces différences en termes de R^2 (variation expliquée) sont négligeables : R^2 de 51,57% pour le rang de monosémie normal, R^2 de 51,6% pour le rang de monosémie de la coupe (*max250*) et finalement R^2 de 49% pour le rang de monosémie de l'exclusion (*under250*).

[158] Notons que certains mots grammaticaux (au niveau des c et des cc) sont des indices désambiguïsateurs intéressants, par exemple *pendant*, qui indique un processus.

Le tableau ci-dessous (Cf. tableau 6.12) visualise toutes les informations sur les longueurs des vecteurs-cc, pour quelques spécificités de l'échantillon des 50 spécificités représentatives. Outre le nombre de c et de cc au seuil de 0,9999, ces informations comprennent le nombre de longueurs des vecteurs-cc, le nombre moyen de vecteurs-cc par longueur et la longueur moyenne par vecteur-cc. Plus les vecteurs-cc sont longs en moyenne (dernière colonne), plus de cc on recense par c.

N°	spécificité	c_0,9999	cc_0,9999	nbr_long_cc-v	moy_ cc-v par long	long_moy_ cc-v
1	*machine*	390	23163	123	3,17	59,39
2	*outil*	282	16050	107	2,64	56,92
3	*permettre*	172	10360	84	2,05	60,23
4	*système*	224	11276	82	2,73	50,34
5	*travail*	146	9679	74	1,97	66,29
6	*Fig*	232	10201	84	2,76	43,97
7	*précision*	107	7197	66	1,62	67,26
8	*arête*	127	7118	65	1,95	56,05
9	*avance*	131	9676	80	1,64	73,86
12	*tour*	173	7480	67	2,58	43,24
15	*m*	197	10131	88	2,24	51,43
19	*abrasif*	89	4166	51	1,75	46,82
20	*Iso*	148	6505	56	2,64	43,95

Tableau 6.12 Extrait de l'échantillon de 50 spécificités : longueur des vecteurs-cc

Comme nous l'avons évoqué ci-dessus, le fait de recenser plus de cc par c en moyenne (Cf. dernière colonne du tableau 6.13) pourra conduire soit à moins de recoupement, si ces cc sont surtout isolés (peu de recoupement avec d'autres cc), soit à plus de recoupement, si ces cc sont surtout partagés (plus de recoupement avec d'autres cc). Nous proposons donc de procéder à l'analyse du recoupement des cc par paire de vecteurs-cc.

6.2.2.2 Le recoupement moyen des cooccurrents des cooccurrents (cc)

Pour étudier le recoupement des cc, nous avons développé un script en Python avec des fonctions déterminant le recoupement par paire de vecteurs-cc (Cf. annexe 9). Les informations pertinentes sont le recoupement total par mot (somme des cc qui figurent dans 2 vecteurs-cc) et le nombre de comparaisons de vecteurs-cc, qui doivent déterminer le recoupement moyen du mot de base. Compte tenu de la longueur des vecteurs-cc, les fonctions en Python permettent de calculer le recoupement relatif moyen, c'est-à-dire le recoupement moyen par spécificité, tout en tenant compte de la longueur des vecteurs-cc. Le recoupement relatif moyen permet de compenser la longueur et ainsi de remédier au problème des vecteurs-cc

extrêmement longs. Si le recoupement (relatif) moyen est élevé, c'est une indication de la monosémie ou de l'homogénéité sémantique du mot de base, étant donné que beaucoup de cc se recoupent et figurent dans deux ou même dans plusieurs vecteurs-cc, ou que beaucoup de cc apparaissent avec plusieurs c.

Les résultats pour l'échantillon des 50 spécificités sont visualisés dans le document en annexe (Cf. annexe 9 : 9.5). En guise de conclusion, les résultats de l'analyse du recoupement moyen des cc et surtout ceux du recoupement relatif moyen des cc confirment l'analyse des 50 spécificités, effectuée à l'aide de notre mesure de recoupement. Les spécificités avec le recoupement moyen le plus élevé (le plus d'homogénéité sémantique) sont notamment *mm/s* et *m/min*. Le recoupement moyen le moins élevé (le plus d'hétérogénéité sémantique) caractérise entre autres *tour* et *abrasif*. Finalement, le recoupement relatif moyen, compensant la longueur des vecteurs-cc responsables du recoupement, aboutit à des résultats plus fiables. Parmi les mots à recoupement relatif moyen élevé, on retrouve *Iso, Fig, etc, mm/s* et *m/min*. Les mots à recoupement relatif moyen plus limité sont *machine, outil, usiner*.

Signalons aussi qu'il semble y avoir une corrélation entre le recoupement moyen et la variation observée dans la longueur des vecteurs-cc, que l'on pourrait analyser à l'aide de l'écart-type des longueurs des vecteurs-cc. Pour ce faire, nous procéderons à une analyse de régression multiple faisant intervenir ces variables de recoupement, ainsi que les variables de fréquence, comme nous le verrons dans la section suivante (Cf. 6.2.3).

6.2.3 La fréquence des cooccurrents des cooccurrents (cc)

La fréquence des cooccurrents des cooccurrents représente le numérateur de la formule de la mesure de recoupement. Cette fréquence nous renseigne sur le nombre de cc isolés et sur le nombre de vecteurs-cc dans lesquels apparaît chaque cc (*cc-type*), pour ainsi dénombrer le nombre d'occurrences (*tokens*) de ce cc-type. Nous procéderons à quelques expérimentations pour l'échantillon de 50 spécificités en matière de fréquence des cc et de pourcentage de cc isolés. A cet effet, nous avons développé un script en Python pour indiquer la fréquence de chaque cc-type et calculer le nombre de cc-types avec une fréquence déterminée. A l'instar de la distribution des longueurs des vecteurs-cc (Cf. 6.2.2.2), on pourra également visualiser la distribution du nombre de cc-types par fréquence. Ces expérimentations nous ont fourni des renseignements particulièrement intéressants, notamment sur les cc isolés (et sur les cc partagés qui en constituent le complément) ainsi que sur la fréquence moyenne par cc-type.

Premièrement, en ce qui concerne les cc isolés, on pourrait avancer l'hypothèse qu'un pourcentage élevé de cc isolés ou non partagés correspond à un faible degré de recoupement et que, inversement, un pourcentage limité de cc isolés correspond à

un degré de recoupement plus important. On s'attendrait donc à ce que les mots intuitivement homogènes affichent des pourcentages de cc isolés plutôt bas et inversement, à ce que les mots hétérogènes se caractérisent par des pourcentages de cc isolés très élevés. Or, les résultats (Cf. annexe 9 : figure A9.2) montrent l'inverse : les mots moins spécifiques ont plus de cc isolés, ce qui paraît contradictoire, à première vue. En fait, ces résultats démontrent qu'il ne faut pas uniquement tenir compte du nombre total de cc et du pourcentage de cc isolés, mais également de la façon dont les cc partagés sont répartis. Par exemple, un cc de fréquence 4 est partagé par 4 c, mais il y a une différence importante entre un cc partagé par 4 c des 6 c au total ou par 4 c des 60 c au total. En effet, pour le degré de recoupement, le dernier pèsera moins lourd, et par conséquent, le nombre total de c est indispensable pour interpréter correctement le recoupement des cc.

Deuxièmement, la fréquence moyenne par cc-type indique par combien de c ce cc (cc-type) est partagé ou donc combien de vecteurs-cc contiennent une occurrence de ce cc-type. Théoriquement, une fréquence moyenne plus élevée signifie que les cc figurent dans plus de vecteurs-cc et qu'ils sont partagés par plus de c, ce qui indique plus de recoupement. Mais sur combien de c au total ? Parmi les 50 spécificités analysées, les mots les plus hétérogènes sémantiquement se caractérisent par la fréquence moyenne la plus élevée, ce qui signifierait en théorie par le recoupement le plus important, car leurs cc sont partagés par plus de c. Toutefois, il faut absolument compenser la fréquence moyenne par le nombre total de c (ou par le nombre total de vecteurs-cc). S'il est vrai que la fréquence moyenne tient compte du nombre de cc au total (*cc-tokens*), elle ne prend aucunement en considération le nombre total de c.

En conclusion des deux dernières sections (6.2.2 et 6.2.3), nous procéderons à une analyse de régression multiple[159] qui fait intervenir les variables pertinentes de recoupement et de fréquence des cc. Celle-ci permettra de rendre compte de la variation du rang de monosémie à partir de plusieurs variables. Nous procédons à cette analyse pour l'échantillon de 50 spécificités, en incluant les facteurs suivants :

- le rang de monosémie
- le nombre moyen de vecteurs-cc par longueur
- la longueur moyenne des vecteurs-cc

---

[159] Les détails techniques d'une analyse statistique de régression multiple seront approfondis dans le chapitre 7 (Cf. 7.2).

- l'écart-type de toutes les longueurs des vecteurs-cc (= variation des longueurs)
- le recoupement relatif moyen
- le pourcentage de cc isolés
- la fréquence moyenne des cc
- l'écart-type de toutes les fréquences des cc (= variation des fréquences)

Les résultats montrent que six facteurs sont significatifs[160] : ensemble ils expliquent 87,9% de la variation du rang de monosémie.

Quatre facteurs se caractérisent par une corrélation négative avec le rang de monosémie : plus les valeurs de ces facteurs seront élevées, plus le rang de monosémie des spécificités est près de 1. D'abord, plus le recoupement relatif moyen est élevé (recouv_rel_moy), plus les 50 spécificités analysées sont monosémiques. Ensuite, plus la longueur moyenne du vecteur-cc est élevée (long_moy_ccv), plus les spécificités sont monosémiques. Puis, on observe de nouveau la corrélation bizarre entre le pourcentage de cc isolés (perc_cc_isol) et l'homogénéité sémantique ou le recoupement. Enfin, plus il y a des cc de fréquences différentes (écart-type des fréquences ou stdev_fq), plus les cc sont partagés et donc plus les mots se recoupent. Toutefois, en ce qui concerne l'interprétation des deux derniers facteurs, les cc isolés et les fréquences différentes, il faudra intégrer aussi le nombre total de c et de cc, ainsi que la façon dont les cc partagés sont répartis, pour calculer correctement le recoupement, comme nous l'avons précisé ci-dessus.

Deux facteurs ont une corrélation positive avec le rang de monosémie des spécificités. Plus le nombre moyen de vecteurs-cc par longueur (moy_ccv_long) est élevé et plus il y a de longueurs différentes de vecteurs-cc (écart-type des longueurs de vecteurs-cc ou stdev_long), plus les mots sont hétérogènes sémantiquement et moins il y a de recoupement.

Il est à noter que cette analyse de régression multiple s'inscrit dans le cadre général des expérimentations visant à mieux raffiner les différents facteurs repris dans la formule de la mesure de recoupement. En plus, les résultats de cette analyse semblent confirmer les résultats des expérimentations antérieures. Nous tenons à insister sur le fait que cette analyse a été effectuée à des fins exploratoires et qu'elle

[160] En raison d'un problème de multicolinéarité (Cf. chapitre 7), le facteur de fréquence moyenne est éliminé de l'analyse de régression multiple (Cf. annexe 9).

ne sert aucunement à tirer des conclusions définitives sur les corrélations éventuelles avec le rang de monosémie. D'ailleurs, elle n'a été conduite que sur un petit échantillon de 50 spécificités, allant des spécificités intuitivement homogènes aux spécificités hétérogènes sémantiquement, des plus aux moins fréquentes et des plus aux moins spécifiques. Par conséquent, notre analyse permet certes de procéder à des observations exploratoires et préliminaires, mais elle ne se prête pas à des conclusions générales. A cette fin, nous procéderons ultérieurement à d'autres analyses de régression multiple qui feront intervenir les différents facteurs de recoupement et de fréquence mais qui seront conduites sur un ensemble plus important de spécificités.

6.2.4 La sensibilité de la mesure de recoupement

Les trois sections précédentes nous ont permis de démontrer l'importance des facteurs qui figurent dans la composition de la mesure de recoupement. Bien évidemment, les facteurs repris dans la formule de la mesure de recoupement soulèvent aussi des questions sur la sensibilité de la mesure, notamment en ce qui concerne le nombre de cooccurrents et la fréquence du mot de base.

La mesure de recoupement s'appuie sur le nombre total de cooccurrents (ou c) d'un mot de base et sur le nombre total de cooccurrents des cooccurrents (ou cc). Il s'agit, rappelons-le, de c et de cc statistiquement très pertinents. La première sensibilité de la mesure de recoupement découle du nombre de cooccurrents et se rapporte au caractère opérationnel de la mesure. D'une part, un mot qui n'a pas de c statistiquement pertinents à ce seuil sévère, n'aura pas de cc non plus, les cc étant rattachés aux c. Comme la mesure est entièrement basée sur les cc et les c, elle n'est pas opérationnelle pour les mots de base n'ayant pas de c au seuil de significativité choisi. La mesure ne génère pas de résultat : 0 dans le numérateur (parce qu'il n'y a pas de cc) et 0 dans le dénominateur (pas de c, pas de cc). D'autre part, un mot avec un seul c au seuil de 0,9999 aura probablement aussi quelques cc significatifs et la mesure pourra générer un résultat. Toutefois, ces cc ne se recouperont jamais, parce qu'il n'y a qu'un seul c et que par conséquent, tous les cc relevés seront différents (non partagés). Mathématiquement, le résultat du calcul de recoupement sera 0 (pas de recoupement du tout), ce qui signifie un résultat hautement hétérogène sémantiquement (hétérogénéité maximale). Mais du point de vue interprétatif, la mesure n'est pas opérationnelle dans le cas d'un seul c significatif, étant donné que le recoupement est techniquement impossible[161].

[161] Notons que, du point de vue logique, on ne peut parler de polysémie qu'à partir d'une étude de deux c différents.

La mesure de recoupement est donc très sensible à un nombre non opérationnel de c (0 c et 1 c). Cette sensibilité ou plutôt cette particularité technique, corollaire des facteurs repris dans la formule de la mesure, impose une restriction importante par rapport aux spécificités pouvant faire l'objet de l'analyse sémantique quantitative automatisée. En effet, les spécificités ayant 0 c ou 1 c seront exclues de nos analyses, pour des simples raisons d'opérationnalité technique.

La deuxième sensibilité de la mesure est liée à la fréquence des mots de base (spécificités). Il est évident qu'un mot aura d'autant plus de chances d'avoir des c statistiquement pertinents qu'il est plus fréquent dans le corpus technique (et par conséquent plus de cc statistiquement pertinents). La fréquence absolue du mot (dans le corpus technique) pourra donc, indirectement, influencer le degré de recoupement de ses cc et donc son rang de monosémie. Mais il y a des exceptions. En effet, plus de c et de cc ne signifient pas toujours moins de recoupement, car il se peut que ces cc supplémentaires soient justement responsables du recoupement. D'ailleurs, plus un mot est fréquent, plus il aura de chances d'apparaître dans les mêmes contextes ou dans des contextes sémantiquement apparentés et plus il aura de chances de se lexicaliser. Cette récurrence, et éventuellement la lexicalisation, donnent lieu à des cooccurrents statistiquement très pertinents et même à des unités polylexicales. Mais, plus un mot est fréquent, plus il aura de chances également d'apparaître dans des contextes plus diversifiés. En effet, dans la langue générale, les mots les plus fréquents sont généralement les plus hétérogènes sémantiquement et se prêtent à la polysémie ou à l'indétermination.

Il convient donc de signaler la sensibilité de notre mesure de recoupement à la fréquence des mots par le biais du nombre de cooccurrents. Les mots les plus fréquents sont susceptibles d'être plus hétérogènes sémantiquement. Toutefois, fréquence ne rime pas toujours avec spécificité et par conséquent, nous maintenons la question principale de cette recherche, à savoir la corrélation entre le rang de monosémie et le rang de spécificité.

Comme nous l'avons évoqué ci-dessus, la mesure de recoupement est sensible à la fréquence et au nombre de c, de par sa nature, mais par contre, elle se révèle insensible à la différence entre l'homonymie, la polysémie et l'indétermination[162]. Si elle permet de distinguer des degrés d'homogénéité sémantique, allant des mots les plus homogènes sémantiquement aux mots les moins homogènes sémantiquement, elle ne discrimine pas puisque tant les homonymes que les mots polysémiques, et

---

[162] On se rappellera que les critères permettant de différencier l'homonymie, la polysémie et l'indétermination ne sont pas toujours efficaces, ni convergents (Cf. chapitre 1).

dans une certaine mesure aussi les mots indéterminés, se caractérisent par l'hétérogénéité sémantique de leurs occurrences (et donc de leurs cooccurrences). Signalons à ce sujet que le but de notre étude n'est pas de distinguer entre ces trois types d'hétérogénéité. Nous développons une mesure pour quantifier la monosémie afin d'automatiser l'analyse sémantique et de la soumettre à des analyses statistiques de régression à grande échelle, c'est-à-dire pour les 4717 spécificités du corpus technique.

6.3 MESURE DE RECOUPEMENT TECHNIQUE

Dans cette dernière partie des mises au point méthodologiques, nous procéderons à l'élaboration d'une mesure de recoupement ou de monosémie technique, en fonction de la spécificité ou de la technicité des cooccurrents des cooccurrents. Cette mesure de recoupement technique pondérée est conçue dans le but de préciser les résultats de la mesure de recoupement et d'aboutir éventuellement à une granularité plus fine.

Contrairement à la première partie sur les expérimentations et analyses permettant de déterminer la configuration la plus stable (Cf. 6.1), nous proposons ici deux variantes de la mesure de recoupement. Notre approche de base (homogénéité sémantique) ainsi que la formule pour la mesure de recoupement se prêtent à des mises au point et permettent en outre d'intégrer d'autres informations afin d'enrichir la mesure de recoupement de base. Pour les analyses statistiques des chapitres suivants (Cf. chapitres 7 et 8), nous proposons dès lors d'adopter les deux mesures de recoupement, à savoir la mesure de recoupement de base et la mesure de recoupement technique pondérée (tenant compte de la technicité des cc). Nous comparons ensuite leurs résultats en termes de rangs de monosémie. Les deux mesures aboutiront à deux analyses de régression, qui pourront être comparées du point de vue de la corrélation respective entre le rang de monosémie (technique) et le rang de spécificité des 4717 spécificités du corpus technique.

Dans la première section, nous expliciterons le principe du recoupement technique (6.3.1). Il mènera à la nouvelle formule de la mesure de recoupement technique, qui sera élaborée dans la deuxième section (6.3.2). Le recoupement technique aboutira finalement à un nouveau calcul de recoupement, dont nous présenterons les premiers résultats pour l'échantillon de 50 spécificités dans la dernière section (6.3.3).

6.3.1 Le principe du recoupement technique

La formule de la mesure de recoupement de base, dont il a été question jusqu'ici (Cf. chapitre 5 et parties 6.1 et 6.2), s'appuie essentiellement sur des informations statistiques de cooccurrence. Cependant, nous aimerions également inclure des informations d'ordre linguistique, notamment la technicité ou la spécificité des cc,

dans le but d'enrichir et de raffiner la mesure de recoupement. Ce sont précisément les cooccurrents des cooccurrents qui sont responsables du recoupement et qui influencent le plus le calcul.

L'idée de base de la nouvelle mesure de recoupement technique repose donc sur la prise en compte de la technicité des cc. Elle repose sur un principe très simple : les cc techniques ou spécifiques du corpus technique pèseront plus lourd sur le recoupement total de tous les cc que les cc généraux (les cc non techniques ou non spécifiques). Lorsqu'un mot de base a plus de cc techniques, responsables du recoupement, ce mot aura un degré de recoupement technique plus élevé. Ainsi, la nouvelle mesure de recoupement technique nous permettra d'évaluer le degré de monosémie technique d'un mot de base ou d'une spécificité de la liste des 4717 spécificités.

Afin de quantifier la technicité ou spécificité des cc, nous proposons de recourir à un facteur de pondération, en fonction de la spécificité des cc. La spécificité des cc est déterminée à partir d'une liste de spécificités (Cf. chapitre 4), c'est-à-dire de toutes les formes graphiques spécifiques du corpus technique. En effet, les cc se situent au niveau des formes fléchies ou formes graphiques.

6.3.2 La formule de la mesure de recoupement technique

La nouvelle mesure de recoupement technique (WLLR) repose sur le LLR (rapport de vraisemblance) pondéré (*weighted LLR*). Elle prend en considération tous les c et tous les cc (dans le dénominateur de la fraction), mais effectue une pondération pour le recoupement des cc (dans le numérateur). Les cc techniques ou spécifiques, ou les cc-clés, se caractérisent par une valeur de LLR importante et par une valeur (1-p) supérieure ou égale à 0,95 (statistiquement significative). Pour établir les facteurs de pondération, nous proposons une nouvelle division de l'échelle des seuils de significativité (1-*p*) des valeurs de LLR.

- Plus le LLR d'un cc est significatif (plus le cc est spécifique ou technique), plus le complément de la valeur p (ou 1-*p*) est élevé et par conséquent, plus ce cc sera important pour le calcul du recoupement technique.

- Moins le LLR d'un cc est significatif (moins le cc est spécifique ou technique), moins le complément de la valeur p (ou 1-*p*) est élevé (mais toujours ≥ 0,95) et dès lors, moins ce cc sera important pour le calcul du recoupement technique.

- Si le cc ne figure pas dans la liste des spécificités, donc s'il n'est pas spécifique du corpus technique, ce cc n'est pas considéré comme technique et il sera très peu important pour le calcul du recoupement technique.

Le facteur de pondération pris en considération pendant le calcul du recoupement sera le WLLR (*weighted LLR*). Le tableau ci-dessous (Cf. tableau 6.13) visualise les différents facteurs de pondération :

complément de la valeur p (ou 1-*p*)	WLLR	
= 1	1	= les cc les plus spécifiques du corpus technique
≥ 0,99	0,9	
≥ 0,985	0,8	
≥ 0,98	0,7	
≥ 0,975	0,6	
≥ 0,97	0,5	
≥ 0,965	0,4	
≥ 0,96	0,3	
≥ 0,955	0,2	= les cc les moins spécifiques du corpus technique
≥ 0,95	0,1	
< 0,95	0,01	= les cc qui ne figurent pas parmi les spécificités

Tableau 6.13 Facteurs de pondération pour la mesure de recoupement technique

Si le cc ne figure pas parmi les spécificités (ou cc-clés), donc si le complément de la valeur p est inférieur à 0,95 (pour p > 0,05), il n'est pas statistiquement significatif. Ce cc non technique sera tout de même pris en considération lors du calcul de recoupement, mais très faiblement, car son poids représente 0,01 et non pas 0 (exclusion du cc). La nouvelle formule de recoupement qui prend en considération le facteur de pondération de la technicité des cc, est explicitée dans la figure 6.13.

$$\sum_{cc} \frac{\text{fq}\, cc \cdot wllr}{\text{nbr total}\, c \cdot \text{nbr total}\, cc}$$

Figure 6.13 Mesure de recoupement technique pondérée

Comme nous l'avons mentionné ci-dessus, un cc plus technique (ou plus spécifique dans le corpus technique) pèsera plus lourd lors du calcul de recoupement. Si le résultat de la mesure de recoupement technique pondérée (WLLR) est élevé, cela signifie que le degré de recoupement technique de ce mot de base est élevé. Il est à noter que généralement, le degré de recoupement technique sera inférieur au degré de recoupement de base. En effet, pour le calcul du recoupement de base, tous les cc sont pris en considération au poids théorique de 1, donc pour un facteur de pondération WLLR théorique de 1 (dans le numérateur de la formule : fq *cc* multiplié par 1).

Pour le calcul du recoupement technique, par contre, seuls les cc les plus techniques auront le poids intégral de 1, les cc un peu moins techniques seront comptabilisés au poids de 0,9 ou de 0,8 et ainsi de suite. Les cc non techniques sont inclus également, mais au poids très faible de 0,01. Si ces cc non techniques se recoupent, leur apport au recoupement total sera limité. Etant donné que la nouvelle mesure de recoupement technique n'exclut aucun cc lors du calcul de recoupement, le dénominateur de la formule reste inchangé. Par conséquent, pour la plupart des spécificités du corpus technique, le degré de recoupement technique sera légèrement moins élevé que le degré de recoupement de base.

De manière générale, plus le degré de recoupement technique est élevé,

- plus il s'approche du degré de recoupement de base
- plus le recoupement se fait par des cc techniques
- plus les cc techniques sont fréquents (et responsables du recoupement)
- plus ces cc techniques seront spécifiques du corpus technique (facteur de pondération plus près de 1, p.ex. 0,9 ou 0,8)

Le fait que le degré de recoupement technique sera légèrement plus limité pour la plupart des spécificités ne veut pas dire que toutes ces spécificités deviennent plus polysémiques du point de vue technique. Les spécificités avec peu de cc techniques qui se recoupent auront simplement un degré de recoupement technique plus limité. Elles seront moins monosémiques « techniquement ». De même, un degré de recoupement technique plus limité ne signifie pas automatiquement un rang de monosémie technique plus bas. Le degré de recoupement technique permet de classer les spécificités analysées par ordre décroissant et dès lors d'accorder un rang de monosémie technique, attribué en fonction du classement par degré. Les spécificités se répartissent donc en rangs en fonction de leur classement. Par conséquent, les rangs de monosémie technique sont susceptibles de subir des changements importants par rapport aux rangs de monosémie de base et ceci en fonction des décalages importants des degrés de recoupement des mots en question.

6.3.3 Premiers résultats : recoupement ou monosémie technique

A l'instar de la mesure de recoupement de base, la nouvelle mesure de recoupement technique pondérée est implémentée dans les scripts en Python. Une première fonction s'appuie sur la liste des formes graphiques spécifiques du corpus technique (Cf. annexe 10), devant aboutir à un dictionnaire Python avec toutes les formes graphiques spécifiques et leur facteur de pondération. Une deuxième fonction calcule pour chaque spécificité son degré de recoupement technique pondéré en

intégrant pour chaque cc de cette spécificité son facteur de pondération, à partir du dictionnaire Python créé antérieurement. Ainsi, pour une liste de plusieurs spécificités, le degré de recoupement technique ou le degré de monosémie technique pourra se calculer automatiquement et générer un document texte avec la spécificité et son degré de recoupement technique à côté.

La mesure de recoupement technique a fait l'objet d'une première expérimentation conduite sur l'échantillon des 50 spécificités représentatives, afin de vérifier son bon fonctionnement. Le tableau ci-dessous (Cf. tableau 6.14) visualise les degrés et les rangs de monosémie et de monosémie technique des 50 spécificités.

N°	spécificité	degré_mono	degré_mono_tech	rang_ v_mono	rang_ v_mono_tech
1	*machine*	0,0231	0,0200	49	49
2	*outil*	0,0240	0,0208	48	48
3	*permettre*	0,0303	0,0258	44	44
4	*système*	0,0280	0,0251	46	46
5	*travail*	0,0307	0,0262	43	42
6	*Fig*	0,0483	0,0403	30	29
7	*précision*	0,0491	0,0451	29	24
8	*arête*	0,0386	0,0339	37	35
9	*avance*	0,0470	0,0426	32	27
10	*usiner*	0,0406	0,0336	36	36
11	*effectuer*	0,0441	0,0348	33	34
12	*tour*	0,0285	0,0255	45	45
13	*puissance*	0,0368	0,0329	39	38
14	*technique*	0,0271	0,0216	47	47
15	*m*	0,0369	0,0305	38	40
16	*table*	0,0321	0,0290	41	41
17	*valeur*	0,0356	0,0312	40	39
18	*etc*	0,0416	0,0336	35	37
19	*abrasif*	0,0313	0,0259	42	43
20	*Iso*	0,0478	0,0397	31	31
21	*électroérosion*	0,0493	0,0400	27	30
22	*tonne*	0,0607	0,0541	22	19
23	*concept*	0,0682	0,0523	21	22
24	*m/min*	0,0718	0,0625	19	17
25	*emboutissage*	0,0496	0,0427	26	26
26	*variable*	0,0537	0,0474	24	23
27	*commander*	0,0556	0,0443	23	25
28	*externe*	0,0491	0,0386	28	32

29	*maîtriser*	0,1224	0,0893	14	16
30	*verre*	0,0535	0,0423	25	28
31	*meulage*	0,0426	0,0365	34	33
32	*fraiser*	0,0888	0,0535	17	20
33	*collaboration*	0,0793	0,0531	18	21
34	*assembler*	0,0715	0,0566	20	18
35	*raboutage*	0,1217	0,1083	15	11
36	*mm/s*	0,1135	0,1019	16	13
37	*insérer*	0,1506	0,1145	11	10
38	*cloison*	0,1415	0,0941	12	15
39	*numériquement*	0,1595	0,1419	10	8
40	*réfrigération*	0,2792	0,1549	5	7
41	*présérie*	0,1282	0,1074	13	12
42	*endommagement*	0,0000	0,0000	--	--
43	*extérieurement*	0,7500	0,6125	1	1
44	*numérotation*	0,2826	0,1907	4	3
45	*réutilisable*	0,2681	0,1746	6	6
46	*microbiologique*	0,1671	0,1391	9	9
47	*puisard*	0,2135	0,1808	8	5
48	*commuter*	0,2500	0,1847	7	4
49	*vidangeur*	0,2917	0,0976	3	14
50	*batch*	0,5137	0,3806	2	2

Tableau 6.14 Echantillon de 50 spécificités : monosémie et monosémie technique

Il est clair que le degré de monosémie technique est partout inférieur au degré de monosémie de base et que les rangs de monosémie et de monosémie technique ne fluctuent pas beaucoup. Toutefois, on observe sur cet échantillon de 50 spécificités que parmi les mots les plus spécifiques et les plus fréquents, certaines spécificités se voient accorder un rang de monosémie technique plus bas que leur rang de monosémie de base. C'est le cas par exemple de *travail, Fig, précision, avance*. Cette différence de rang indique que ces mots se caractérisent par une homogénéité sémantique technique plus grande que leur homogénéité sémantique de base (générale). Si ces mots sont polysémiques, leur polysémie est donc plutôt générale (Cf. chapitre 7). Ces modifications de rang de monosémie technique s'accompagnent bien sûr d'autres modifications dans le sens inverse. En effet, d'autres spécificités de cette liste se voient accorder un rang de monosémie technique un peu plus élevé : elles ont une homogénéité sémantique technique moins grande que leur homogénéité sémantique de base (générale). C'est le cas par exemple de *collaboration, maîtriser*, etc.

Rappelons que ces observations exploratoires cadrent dans les expérimentations et les mises au point visant à vérifier les différents facteurs de la formule de la mesure de recoupement technique. Il va de soi que des expérimentations sur des échantillons plus larges et surtout sur la liste entière des 4717 spécificités permettront d'aboutir à des résultats plus concluants, plus fiables et mieux interprétables du point de vue linguistique.

En plus, les données quantitatives de rang de monosémie et de rang de monosémie technique, appliquées à toutes les spécificités, se prêteront à des analyses statistiques de régression, qui permettront d'évaluer la corrélation entre le rang de monosémie et le rang de spécificité, ainsi que la corrélation entre le rang de monosémie technique et le rang de spécificité. Ces analyses de régression simple et multiple et leurs résultats feront l'objet du chapitre suivant (Cf. chapitre 7).

PARTIE III

Résultats et interprétations

Chapitre 7

Analyses de régression de base

Cette étude se termine par les résultats des analyses statistiques de régression et par l'interprétation des résultats. La double approche méthodologique, à savoir l'analyse des spécificités (Cf. chapitre 4) et l'analyse des cooccurrences (Cf. chapitre 5), aboutit à des données quantitatives de spécificité et d'homogénéité sémantique. Ces dernières feront l'objet d'analyses statistiques de régression, qui mettent en évidence leur corrélation.

Dans ce chapitre, nous procéderons donc à des analyses de régression de base, c'est-à-dire pour la liste des 4717 spécificités du corpus technique. Dans un premier temps, une analyse statistique de régression simple permettra d'évaluer l'impact du rang de spécificité sur le rang de monosémie. Ainsi, les résultats de l'analyse de régression simple fourniront la réponse à la question principale de notre recherche, comme nous verrons dans la première partie du chapitre (7.1). Etant donné que la monosémie ou l'homogénéité sémantique n'est pas uniquement influencée par la spécificité, nous procéderons également à une analyse statistique de régression multiple, qui sera décrite dans la deuxième partie (7.2). L'analyse de régression multiple fera intervenir plusieurs variables indépendantes susceptibles d'influer sur le rang de monosémie du mot de base (mot spécifique), entre autres le rang de spécificité de ce mot de base, sa fréquence, sa classe lexicale et le nombre de classes lexicales.

Nous discuterons les résultats des analyses de régression simple et multiple pour les 4717 spécificités et nous tenterons en ce faisant de trouver une solution linguistique aux problèmes techniques posés par les analyses statistiques de régression.

7.1 ANALYSE DE RÉGRESSION SIMPLE

Une analyse de régression simple vise à étudier l'impact d'une variable indépendante ou variable explicative sur une deuxième variable, la variable

dépendante ou variable expliquée (ou encore variable à expliquer)[163]. Dans notre étude, la variable indépendante (VI) est le rang de spécificité ; la variable dépendante (VD) est le rang de monosémie. Le résultat d'une analyse de régression simple est le pourcentage de variation expliquée R^2, appelé aussi le coefficient de détermination. Ce pourcentage représente le pourcentage de la variation du rang de monosémie que l'on pourra expliquer ou prédire à partir de la variation du rang de spécificité d'un ensemble de données, en l'occurrence la liste des 4717 spécificités. Le résultat R^2 de l'analyse de régression comprend toujours une valeur p, indiquant la significativité statistique du modèle de régression et donc la fiabilité de la capacité prédictive du modèle.

Dans la première section de cette partie, nous commenterons les résultats des analyses statistiques[164] pour les 4717 spécificités (7.1.1). Ensuite, nous présenterons les résultats pour le rang de monosémie technique (7.1.2). La troisième section sera consacrée au problème de l'hétéroscédasticité (7.1.3), que nous résoudrons dans la section suivante (7.1.4), en adoptant non seulement des solutions techniques, mais également des solutions compatibles avec l'interprétation linguistique des données. Finalement, nous essaierons de caractériser les spécificités plutôt générales (7.1.5) et de formuler une conclusion (7.1.6).

7.1.1 Résultats de l'analyse de régression simple

7.1.1.1 Deux variables : le rang de spécificité et le rang de monosémie

Précisons d'abord la dénomination adoptée pour le rang de spécificité et le rang de monosémie. Les données quantitatives se présentent sous forme de degrés. Le classement des 4717 spécificités, en fonction de leur degré de spécificité, permet d'accorder un rang de spécificité. Un rang de spécificité proche de 1 caractérise les mots les plus spécifiques, un rang proche de 4717 indique les mots les moins spécifiques de la liste. Plus le rang est élevé, moins le mot est spécifique du corpus technique. Les mots avec un degré de spécificité identique, c'est-à-dire avec une valeur de LLR identique, auront le même rang de spécificité. Les nouveaux rangs de spécificité, exprimés par la variable rang_v_spec[165], permettent une comparaison

[163] Les équivalents *variable prédictive* (variable indépendante) et *variable prédite* (variable dépendante) mettent en évidence le caractère prédictif du modèle de régression.

[164] Les analyses statistiques de régression et de corrélation, ainsi que les visualisations sont réalisées à l'aide du logiciel statistique R : http://www.r-project.org.

[165] Nous optons pour la dénomination *rang_v_spec* (rang de spécificité identique pour des valeurs de LLR identiques), contrairement à *rang_spec* (indiquant la numérotation des spécificités, de 1 à 4717, sans tenir compte de valeurs identiques, donc sans rangs identiques).

plus juste des spécificités ayant la même valeur de LLR ou le même degré de spécificité. Parmi les rangs de la variable rang_v_spec, le même rang pourra figurer plusieurs fois. Les rangs de spécificité ci-dessous, rang_v_spec 195, 196, 196, 198, 199 … montrent par exemple que *formage* et *taraudage* ont un LLR ou un degré de spécificité identique (Cf. tableau 7.1). Le mot suivant aura le rang 198, étant donné que le rang 197 n'a pas été accordé et que la numérotation habituelle des rangs reprend lorsque le degré de spécificité est de nouveau différent[166].

rang_spec	LLR	**rang_v_spec**	rang_v_mono_0,9999	lemme
195	1345,54038	195	4124	*lubrification*
196	**1341,85223**	**196**	4428	*formage*
197	**1341,85223**	**196**	4390	*taraudage*
198	1334,23831	198	4064	*bille*
199	1334,09453	199	4427	*etc*

Tableau 7.1 Rangs et degrés de spécificité identiques (LLR) : rang_v_spec

7.1.1.2 Corrélation négative et variation expliquée

Avant de discuter les résultats, il est intéressant d'étudier la corrélation entre les deux variables à l'aide du coefficient de corrélation Pearson, parce qu'il donne une première indication de la relation entre les deux variables (Cf. tableau 7.2). Le coefficient de corrélation Pearson (-0,72) montre une corrélation négative entre le rang de spécificité et le rang de monosémie. Par conséquent, les mots les plus spécifiques du corpus technique ne sont pas les plus monosémiques, au contraire.

```
        Pearson's product-moment correlation

data:  rang_v_mono_0.9999 and rang_v_spec
t = -70.8669, df = 4715, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7317133 -0.7040630
sample estimates:
       cor
-0.7181715
```

Tableau 7.2 Corrélation : rang de monosémie ~ rang de spécificité

[166] La même dénomination de *rang_v_...* sera adoptée pour les rangs de monosémie (au seuil de significativité de 0.9999), à savoir *rang_v_mono_0.9999*. Ultérieurement, il en va de même pour les rangs de fréquence technique et pour les rangs de fréquence générale, respectivement *rang_v_freq1* et *rang_v_freq2*. Les fonctions d'Excel permettent de réaliser facilement ces opérations de numérotation de rangs identiques.

Pour mieux étudier les détails de cette corrélation négative, nous recourons à une analyse statistique de régression linéaire simple, dans laquelle les deux variables sont considérées comme les deux axes d'une visualisation en deux dimensions. En abscisse (axe X), on trouve la variable indépendante, soit le rang de spécificité. En ordonnée (axe Y), on observe la variable dépendante ou le rang de monosémie. Les 4717 spécificités se caractérisent par des valeurs pour chacune de ces deux variables et se prêtent dès lors à une visualisation en fonction des deux axes, sous forme d'un « nuage de points », qui visualise les valeurs observées pour les deux variables (Cf. figure 7.1).

Le fait de prédire une variable à partir d'une deuxième variable correspond à un modèle de régression linéaire. Ce modèle a pour objectif de faire passer une ligne droite[167] à travers le nuage de points et de visualiser ainsi la valeur prédite ou estimée par le modèle de régression pour chaque point. Toutefois, les points des valeurs observées (valeurs originales) ne se situent que très rarement sur une droite. En effet, généralement, le rapport entre les deux variables n'est pas parfaitement linéaire. Comme la prédiction des valeurs de la variable dépendante en fonction de la droite de régression signifie une perte d'informations, on essaie de limiter celle-ci en cherchant la droite qui corresponde le mieux aux valeurs observées (*best fit*) et qui minimise la différence entre les valeurs observées et les valeurs estimées. Cette droite est appelée la « droite des moindres carrés » : elle minimise la distance entre chaque point et la droite. Nous recourons donc à la droite de régression ou droite des moindres carrés pour prédire de nouvelles valeurs du rang de monosémie (axe Y) à partir des valeurs du rang de spécificité (axe X).

Pour chaque point, la différence entre la valeur observée (Y) et la valeur estimée (Y') (située sur la droite de régression) est appelée le résidu (la valeur résiduelle) ou l'erreur (ε), car elle correspond à l'erreur qu'on commet en prédisant la valeur de la variable dépendante (Y) à partir des valeurs estimées (Y') données par la droite de régression. Les résultats de l'analyse statistique de régression linéaire simple pour les 4717 spécificités sont visualisés dans le tableau ci-dessous (Cf. tableau 7.3). Cette analyse de régression est hautement significative (valeur $p < 2.2e^{-16}$) et le coefficient de détermination R^2 est de 0,5157. Notons que le coefficient de

---

[167] Formule de la droite de régression : $Y' = a + bX$.

X est la variable indépendante ; Y' est la valeur estimée de la variable dépendante ; a est l'ordonnée à l'origine (*intercept*) du modèle, donc la valeur de Y' lorsque X=0 ; b est le coefficient de régression ou la pente, c'est-à-dire la variation de Y' pour une variation d'une unité de X. Les valeurs de a et b pour les 4717 spécificités sont données par le modèle de régression (Cf. tableau 7.3).

détermination R² du modèle de régression équivaut au carré du coefficient de corrélation R (-0,7181). Comme le coefficient de détermination R² mesure la quantité de variation expliquée par la droite de régression par rapport à la variation totale, il correspond au pourcentage de variation expliquée. La variation du rang de spécificité permet donc d'expliquer 51,57% de la variation du rang de monosémie.

```
Call:
lm(formula = rang_v_mono_0.9999 ~ rang_v_spec, data = m)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4066.70091   27.79260  146.32   <2e-16 ***
rang_v_spec   -0.73239    0.01033  -70.87   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 949 on 4715 degrees of freedom
Multiple R-Squared: 0.5158,      Adjusted R-squared: 0.5157
F-statistic:  5022 on 1 and 4715 DF,  p-value: < 2.2e-16
```

Tableau 7.3 Régression simple : rang de monosémie ~ rang de spécificité

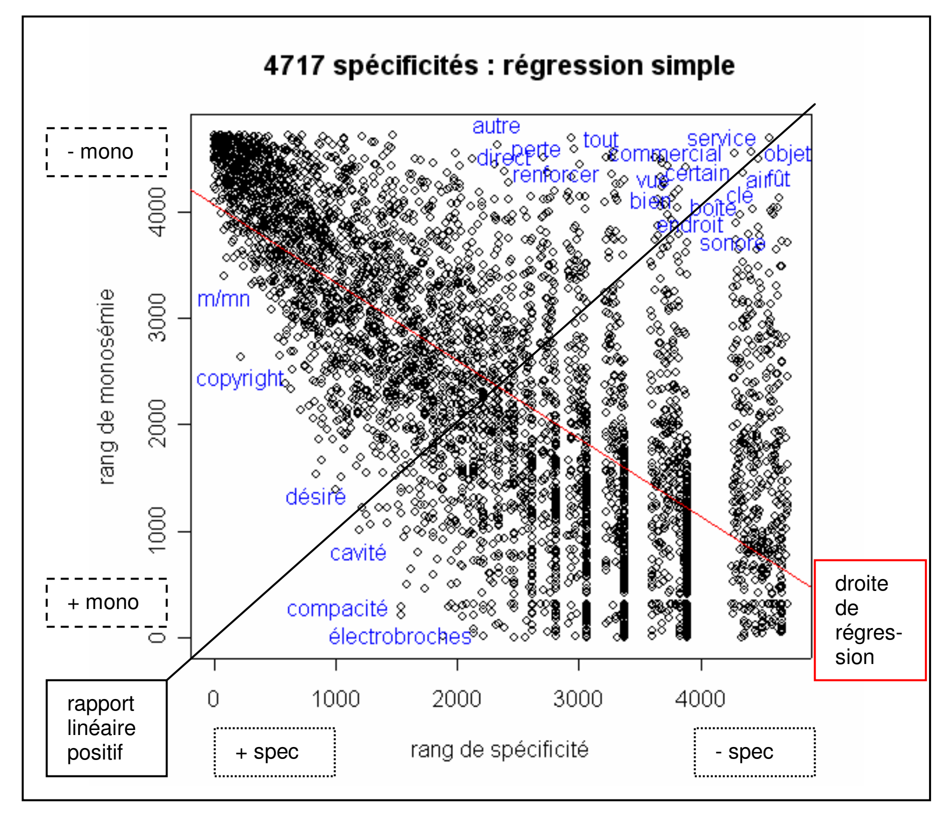


Figure 7.1 Régression simple : rang de monosémie ~ rang de spécificité

La figure 7.1 ci-dessus visualise les valeurs observées des 4717 spécificités, en fonction de leur rang de spécificité (axe X) et en fonction de leur rang de monosémie (axe Y). Les valeurs observées (*Y*) sont marquées par des points noirs et constituent le nuage de points à travers duquel on fait passer la droite de régression (indiquée en rouge). La droite de régression visualise les valeurs estimées (*Y'*) pour le rang de monosémie (variable dépendante), pour chaque rang de spécificité (variable indépendante). Comme la droite de régression est inclinée vers le bas, elle indique une corrélation négative entre le rang de spécificité et le rang de monosémie des 4717 spécificités.

7.1.1.3 Interprétation linguistique globale

La visualisation de la régression linéaire simple ci-dessus (Cf. figure 7.1) montre que les mots les plus spécifiques, visualisés à gauche, se trouvent généralement en haut de la représentation. Dès lors, les mots les plus spécifiques du corpus technique sont les moins monosémiques ou les moins homogènes sémantiquement, par exemple *machine, pièce, tour*. Par contre, les mots les moins spécifiques à droite de la représentation se situent majoritairement en bas et sont donc plutôt monosémiques (*rationnellement, télédiagnostic, autosurveillance*). La visualisation par la droite de régression descendante indique clairement le rapport négatif entre les deux et va donc à l'encontre d'un rapport linéaire positif que l'on pourrait attendre si la thèse des monosémistes traditionnels se vérifiait. Notons d'emblée que la figure 7.1 soulève la question de la pertinence de la régression linéaire, qui sera abordée ci-dessous (Cf. 7.1.4.1), étant donné que le rapport entre les deux variables ne semble pas tout à fait linéaire.

Les résultats de l'analyse de régression simple ainsi que la visualisation permettent donc d'infirmer la thèse traditionnelle. En effet, les mots les plus spécifiques de notre corpus technique d'analyse sont les plus polysémiques. Inversement, les mots les moins spécifiques s'avèrent les plus monosémiques et cela à quelques exceptions près, notamment *service, objet, commercial,* etc. qui se situent en haut à droite, assez loin de la droite de régression. Ces mots sont très peu spécifiques et très peu monosémiques. Citons également quelques exceptions à gauche en bas, telles que *électrobroches* et *cavité* (Cf. figure 7.1), des mots assez spécifiques et plutôt monosémiques. Ces exceptions à la tendance générale seront discutées ci-dessous (Cf. 7.1.3 et 7.1.4).

7.1.2 Le rang de monosémie technique

Dans le but de préciser les résultats de la mesure de monosémie de base, nous avons élaboré une mesure de recoupement ou de monosémie technique (Cf. chapitre 6). Cette mesure tient compte, dans le calcul du recoupement des cooccurrents des cooccurrents (cc), de la spécificité des cc dans le corpus technique, donc de la

technicité des cc. Plus les cc sont spécifiques du corpus technique, plus ils pèsent lourd dans le calcul du recoupement. La mesure de monosémie technique permet ainsi de déterminer dans quelle mesure le mot de base se caractérise par l'homogénéité sémantique technique.

Le classement des spécificités (ou mots de base) par ordre décroissant de degré de monosémie technique permet d'attribuer un rang, qui soulève deux questions intéressantes. Nous nous demandons si les spécificités se caractérisent par un comportement semblable ou différent par rapport au rang de monosémie de base et par rapport au rang de monosémie technique. En plus, le rang de monosémie technique, pourrait-il conduire à une distinction opérationnelle entre la monosémie technique et la monosémie générale des spécificités analysées ?

7.1.2.1 Particularités du rang de monosémie technique

La mesure de recoupement ou de monosémie technique prévoit une pondération en fonction de la spécificité (LLR) des cc. Par rapport à la mesure de monosémie de base, le dénominateur de la fraction (Cf. figures 6.11 et 6.13) reste inchangé (nombre total de c × nombre total de cc). Par contre, le numérateur est affecté par le facteur de pondération *wllr* (fq *cc* × *wllr*). Dès lors, plus la fréquence du cc (*token*) est élevée, plus l'impact du facteur de pondération sera important, étant donné que sommer sur tous les cc revient à considérer par exemple 6 fois la fréquence 6 (= 6^2). Rappelons qu'un cc de fréquence 6 figure avec 6 c. En effet, si la fréquence du cc est plus élevée, le cc est plus partagé et, par voie de conséquence, il pèsera plus lourd sur le recoupement et subira plus fortement l'impact du facteur de pondération.

La comparaison croisée de la fréquence d'un cc et de son facteur de pondération (*wllr*), qui reflète sa spécificité dans le corpus technique, permet de distinguer quatre cas de figure décrits ci-dessous (Cf. tableau 7.4). Les cc spécifiques du corpus technique se voient attribuer les facteurs de pondération les plus élevés (au maximum 1). Les cc généraux se caractérisent par le facteur de pondération minimal de 0,1 (Cf. chapitre 6).

fréquence du cc	facteur de pondération (wllr) du cc	contribution au degré de mono tech	contribution au degré de mono	conclusion
élevée (p.ex. 6)	élevé (1 ou 0,9)	$6^2 \times 0{,}9 = 32{,}4$	$6^2 = 36$	mono tech.
minimale (p.ex. 1)	élevé (1 ou 0,9)	$1^2 \times 0{,}9 = 0{,}9$	$1^2 = 1$	poly tech.
minimale (p.ex. 1)	limité (0,1)	$1^2 \times 0{,}1 = 0{,}1$	$1^2 = 1$	poly gén.
élevée (p.ex. 6)	limité (0,1)	$6^2 \times 0{,}1 = 3{,}6$	$6^2 = 36$	mono gén.

Tableau 7.4 Comparaison croisée : fréquence et spécificité du cc

1) Si la fréquence du cc est plutôt élevée (plus de recoupement) et si le cc est plutôt technique ou spécifique (*wllr* de 1 ou 0,9), sa contribution au degré de monosémie technique sera importante. Un mot de base avec beaucoup de cc techniques fréquents se caractérisera par un degré de monosémie technique très élevé et dès lors par un rang de monosémie technique plutôt bas (ou monosémique).

2) Si la fréquence du cc est minimale (pas de recoupement) et si le cc est plutôt technique (*wllr* de 1 ou 0,9), sa contribution au degré de monosémie technique sera faible. Dans ce cas de figure, le degré de monosémie technique sera faible globalement et conduira à un rang de monosémie technique plutôt élevé (ou polysémique).

3) Si la fréquence du cc est minimale (pas de recoupement) et si le cc est général (*wllr* de 0,1), sa contribution au degré de monosémie technique sera bas à l'extrême. Par conséquent, le degré de monosémie technique sera très faible et conduira à un rang de monosémie technique encore plus élevé.

4) Si la fréquence du cc est plutôt élevée (plus de recoupement) et si le cc est général (*wllr* de 0,1), sa contribution au degré de monosémie technique sera très limitée, en dépit de sa fréquence importante. En plus, dans ce cas de figure, le facteur de pondération très faible de 0,1 génère la différence la plus grande possible entre le degré de monosémie et le degré de monosémie technique. Si un mot de base a beaucoup de cc généraux fréquents, son degré de monosémie technique sera beaucoup plus faible que son degré de monosémie. Le rang de monosémie technique sera plutôt élevé.

Bien évidemment, les cc d'un mot de base ne se situent pas tous dans le même cas de figure (Cf. tableau 7.4) et l'analyse n'est pas aussi aisée qu'elle ne paraît. Toutefois, les caractéristiques des cc donnent une indication fiable du type de monosémie du mot de base. Si les cc d'un mot de base sont majoritairement des cc techniques, (très) spécifiques du corpus technique, et s'ils sont plutôt fréquents (et donc responsables de recoupement), le mot de base se caractérise par la monosémie technique.

Si, en revanche, les cc d'un mot de base se situent principalement dans un des autres cas de figure, le calcul de la mesure de monosémie technique conduira à un rang de monosémie technique plutôt élevé, c'est-à-dire polysémique ou même très polysémique (pour un degré de monosémie technique plutôt bas ou même très bas). Toutefois, l'explication des cas de figure ci-dessus montre qu'un tel résultat ne coïncide pas toujours avec la polysémie technique. D'autres variables seront nécessaires pour déterminer si les cc sont majoritairement généraux et si, par

conséquent, un degré de monosémie technique plutôt bas ou même très bas cache respectivement de la monosémie générale ou de la polysémie générale.

Nous procéderons plus loin à des expérimentations basées sur des variables supplémentaires des cc, notamment la fréquence moyenne des cc et leur technicité moyenne (Cf. 7.1.4), pour caractériser le type de monosémie en fonction du type de mot de base.

7.1.2.2 Corrélation négative et variation expliquée

Le rang de monosémie technique se prête aussi à des analyses de corrélation et de régression simple, qui permettent de déterminer si et dans quelle mesure le rang de spécificité d'un mot de base explique ou prédit son rang de monosémie technique. Le coefficient de corrélation (-0,65) est moins élevé que le coefficient de corrélation entre le rang de spécificité et le rang de monosémie de base (-0,72) (Cf. tableau 7.2).

L'analyse de régression simple pour le rang de monosémie technique aboutit à des résultats similaires : elle est hautement significative ($p < 2.2e^{-16}$) et le pourcentage de variation expliquée R^2 est de 42,74%. La variation du rang de spécificité permet donc d'expliquer 42,74% de la variation du rang de monosémie technique, tandis qu'elle explique 51,57% de la variation du rang de monosémie de base (Cf. tableau 7.3). Il s'ensuit que, pour les 4717 spécificités, le rang de spécificité est une variable explicative ou prédictive moins bonne et moins fiable pour le rang de monosémie technique que pour le rang de monosémie.

Les deux variables, à savoir le rang de spécificité et le rang de monosémie technique, sont visualisées ci-dessous (Cf. figure 7.2). La droite de régression indiquée en rouge (ligne continue) s'incline vers le bas et visualise donc la tendance négative, bien qu'elle soit moins claire que la tendance négative visualisée dans la figure précédente du rang de monosémie (Cf. figure 7.1). Dans cette figure 7.2, nous avons superposé, en tireté, la droite de régression de la figure précédente. La comparaison de ces deux droites de régression indique que la droite de régression du rang de monosémie technique (ligne continue) descend un peu pour les spécificités les plus spécifiques du corpus technique (rangs < 1000). En plus, elle augmente un peu pour les spécificités les moins spécifiques du corpus technique (rangs > 3000). Cette différence de position de la droite de régression technique (ligne continue) s'explique par la position des 4717 spécificités, et plus particulièrement par leur rang de monosémie technique, qui est la variable dépendante, étant donné que la variable indépendante (le rang de spécificité) reste inchangée.

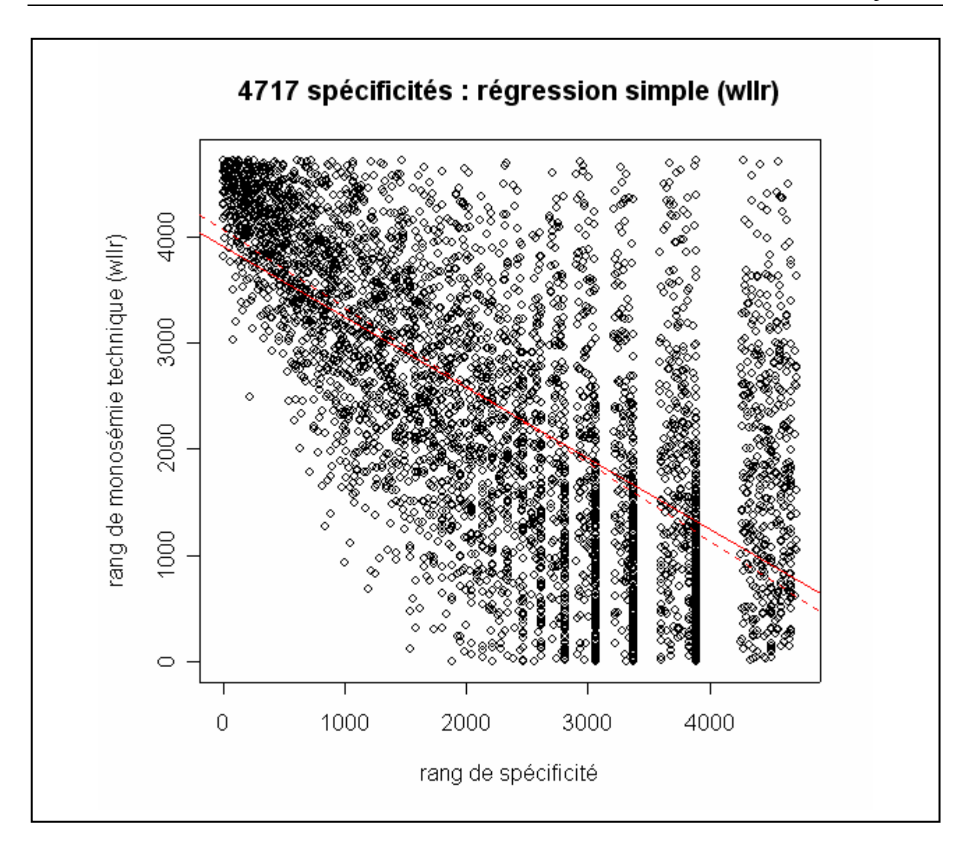


Figure 7.2 Régression simple : rang de monosémie technique ~ rang de spécificité

Grosso modo, le déplacement de la droite de régression pour le rang de monosémie technique par rapport au rang de monosémie de base signifie que, d'une part, (la plupart) des mots plus spécifiques ont tendance à devenir un peu plus monosémiques, si l'on considère principalement leurs cc techniques. D'autre part, (la plupart) des mots moins spécifiques ont tendance à devenir un peu plus polysémiques, si l'on considère principalement leurs cc techniques (Cf. formule de la mesure de recoupement technique).

D'ailleurs, la comparaison détaillée du nuage de points dans les deux figures (Cf. figures 7.1 et 7.2) révèle clairement que dans la figure 7.2, le nuage de points est plus dispersé que dans la figure 7.1. Dans la figure 7.2, un certain nombre de mots plus spécifiques s'orientent davantage vers le coin inférieur gauche de la représentation (plus monosémiques) et un certain nombre de mots moins spécifiques s'orientent davantage vers le coin supérieur droit (plus polysémiques). Comme la tendance linéaire est moins claire, la question de la pertinence de la régression linéaire se pose également, et a fortiori, pour cette analyse qui fait intervenir le rang de monosémie technique.

7.1.2.3 Interprétation linguistique globale

Généralement, en fonction de la mesure de recoupement technique pondérée et donc du rang de monosémie technique, les mots les plus spécifiques du corpus technique, à gauche de la visualisation, sont les moins monosémiques. Comme les mots les moins spécifiques se situent principalement en bas de la visualisation, ils sont plutôt monosémiques, à quelques exceptions près. Toutefois, les résultats pour le rang de monosémie technique sont moins concluants que ceux obtenus pour le rang de monosémie de base.

En plus, la comparaison des résultats visualise un faible déplacement de la droite de régression : un peu plus monosémique à gauche pour les mots plus spécifiques et un peu plus polysémique à droite pour les mots moins spécifiques. Ce léger déplacement de la droite de régression pourrait s'interpréter comme un léger effet de la thèse monosémiste. En effet, les mots les plus spécifiques, ayant probablement beaucoup de cc techniques, sont un peu plus monosémiques techniquement. Néanmoins, la tendance est toujours négative : les mots les plus spécifiques se situent toujours du côté des rangs les moins monosémiques. Par conséquent, la tendance observée pour le rang de monosémie technique s'oppose aussi à la corrélation positive préconisée par les monosémistes, tout comme la tendance pour le rang de monosémie.

Finalement, il convient d'étudier les particularités de la mesure de monosémie technique (Cf. 7.1.2.1), afin d'aboutir à une interprétation linguistique plus nuancée. Si l'on tient compte de la technicité des cc, les mots les plus spécifiques, ayant probablement le plus de cc spécifiques ou techniques, deviennent plus monosémiques. Or, ce qui joue un rôle important dans la mesure de recoupement technique, ce n'est pas uniquement le pourcentage de cc techniques ou spécifiques par rapport au nombre total de cc, mais également et surtout le degré de spécificité des cc (donc l'impact du facteur de pondération) et leur fréquence ou la mesure dans laquelle les cc techniques se recoupent (Cf. tableau 7.4 ci-dessus). Plus les cc sont techniques et plus ils sont fréquents, plus le degré de monosémie technique sera élevé. Donc, les rangs de monosémie technique un peu plus monosémiques pour les mots spécifiques s'expliquent principalement par le fait que leurs cc techniques pèsent beaucoup plus lourd dans la formule et sont ainsi responsables des résultats plus élevés du degré de monosémie technique. Le fait que les mots spécifiques deviennent un peu plus monosémiques techniquement, plutôt que de corroborer la thèse monosémiste, découle de la pondération de la formule de recoupement technique et du recoupement des cc techniques, bien qu'il soit plutôt faible. Les détails de la fréquence moyenne des cc (recoupement) et de leur technicité moyenne seront élaborés ultérieurement (Cf. 7.1.5.2 et 7.1.5.3) et cela pour différents sous-ensembles des 4717 spécificités.

7.1.3 Le problème de l'hétéroscédasticité

L'application d'une analyse de régression linéaire simple impose certaines conditions : des observations indépendantes, un rapport linéaire entre les variables *X* et *Y* et, en plus, l'homogénéité et la normalité des erreurs ou des résidus (distribués normalement autour de zéro), c'est-à-dire leur homoscédasticité. Le problème de l'hétéroscédasticité indique que les variances des erreurs ne sont pas constantes. Certaines observations ont en effet des résidus très importants et se situent très loin de la droite de régression des valeurs estimées.

En cas d'hétéroscédasticité, les erreurs standard des estimations du modèle de régression simple (Cf. tableau 7.3 : *Std. error*) sont incorrectes et souvent sous-estimées. Par conséquent, les inférences statistiques du modèle sont invalides et les intervalles de confiance ainsi que les tests basés sur les erreurs standard sont incorrects. L'hétéroscédasticité signifie donc que les estimateurs de la méthode des moindres carrés ne sont pas efficaces et que la droite de régression n'est pas la meilleure prédiction possible.

Or comment détecter l'hétéroscédasticité ? La visualisation des résultats de l'analyse de régression simple (Cf. figure 7.1) montre que certains mots se situent (très) loin de la droite de régression, par exemple *service, objet, commercial*. Ces mots se caractérisent par une distance très importante entre la valeur observée (rang de monosémie près de 4700) et la valeur du rang de monosémie estimée par la droite de régression (environ 500) et donc par une erreur d'estimation très importante. On observe également des mots en dessous de la droite de régression, dans la partie inférieure gauche de la représentation : des mots plus spécifiques à résidus négatifs. Toutefois, on observe moins de mots à résidus négatifs qu'il y a de mots à résidus positifs dans la partie supérieure droite de la représentation.

La visualisation des résidus ci-dessous (Cf. figure 7.3) soulève également la question de l'hétéroscédasticité, les résidus étant indiqués en ordonnée (axe Y), en fonction des valeurs estimées du rang de monosémie (axe X). Normalement, une régression linéaire simple se caractérise par l'homoscédasticité des résidus : les résidus suivent une distribution normale, ils sont répartis de façon aléatoire et homogène autour de zéro (autant de résidus positifs que négatifs) ou autour de la droite de régression, sans qu'il y ait de patron. A gauche, pour des valeurs faibles de rang de monosémie estimé, on observe les mots à résidus positifs importants, tels que *service, commercial, objet*. Ce sont des mots peu spécifiques et peu monosémiques, qui se caractérisent par une erreur d'estimation très importante. Ces mots sont plus polysémiques qu'on n'aurait cru en tenant compte de leur rang de spécificité. A droite de la visualisation, les mots à résidus négatifs représentent les

mots plus monosémiques que prévu à partir de leur rang de spécificité, par exemple *électrobroches* et *cavité*.

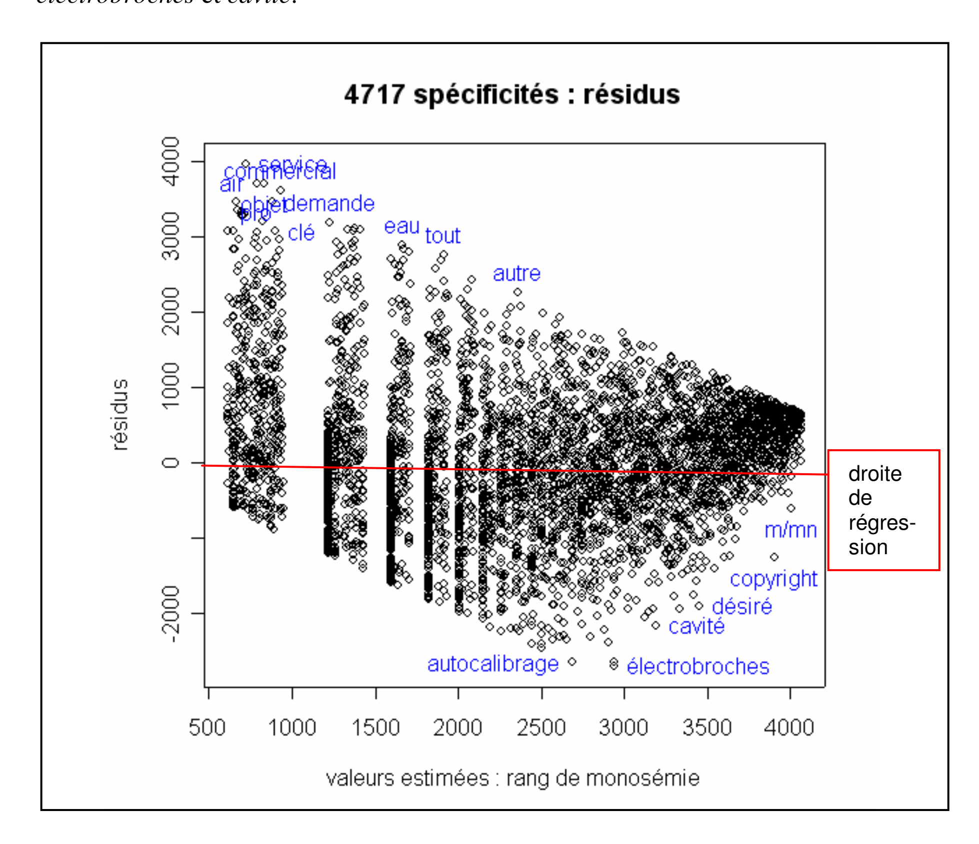


Figure 7.3 Régression simple : visualisation des résidus

Le problème technique de l'hétéroscédasticité peut être saisi à l'aide du test statistique de Goldfeld-Quandt (gqtest), implémenté dans le logiciel R. Si la valeur p du gqtest est statistiquement significative (< 0,05), l'hypothèse nulle d'homoscédasticité (ou de variances constantes) est rejetée et l'hétéroscédasticité est détectée. Les résultats du test statistique de Goldfeld-Quandt, visualisés ci-dessous (Cf. tableau 7.5), confirment donc les observations des visualisations précédentes (Cf. figures 7.1 et 7.3).

```
        Goldfeld-Quandt test

data:  rang_v_mono_0.9999 ~ rang_v_spec
GQ = 2.0725, df1 = 2357, df2 = 2356, p-value < 2.2e-16
```

Tableau 7.5 Gqtest : hétéroscédasticité

7.1.3.1 Exploration des mots à résidus importants

Afin de mieux comprendre les particularités des mots ou des spécificités à résidus importants, responsables du problème de l'hétéroscédasticité, nous procédons à une première analyse exploratoire. Rappelons que les mots à résidus positifs importants sont moins monosémiques que prévu en fonction de leur rang de spécificité. Les mots à résidus négatifs, par contre, sont plus monosémiques que prévu.

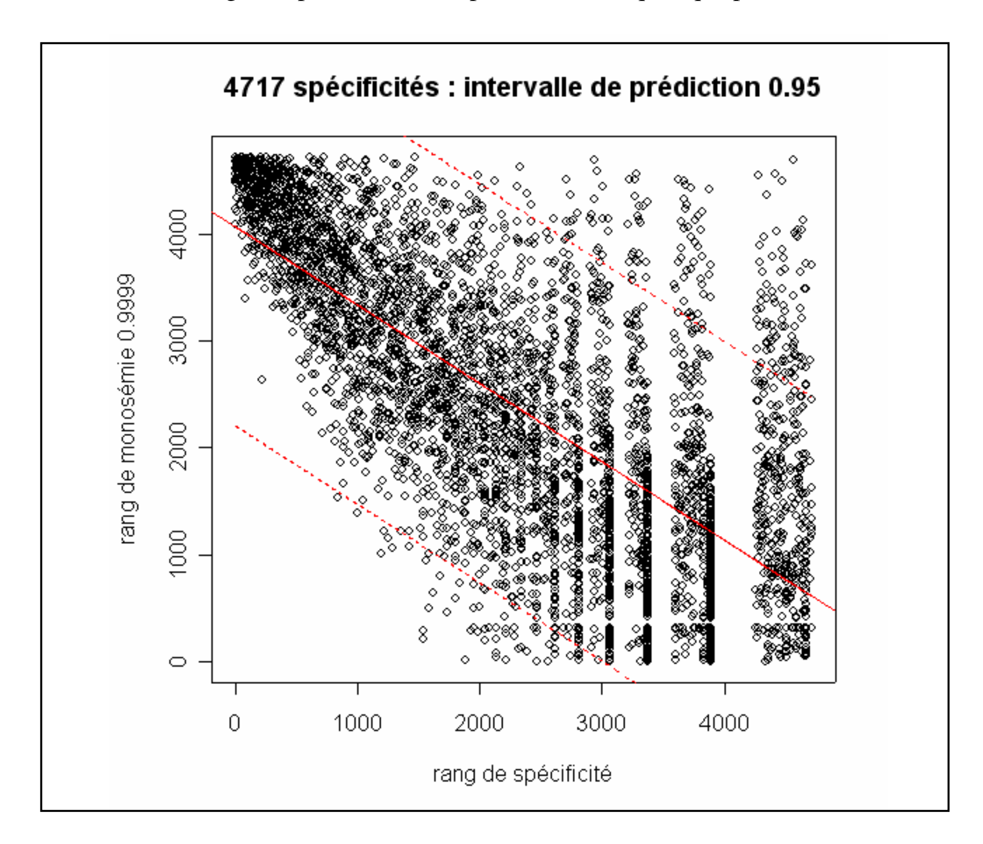


Figure 7.4 Régression simple : intervalle de confiance (prédiction)

La visualisation ci-dessus (Cf. figure 7.4) montre l'intervalle de confiance à 0,95 pour la prédiction individuelle des valeurs du rang de monosémie. Les mots à résidus importants se situent en dehors des deux bandes de prédiction en pointillé et donc en dehors de l'intervalle de confiance (à une distance d'environ 2000 de la droite de régression). Pour ces mots dont la plupart se situent à droite en haut, la déviation n'est plus due au hasard : il s'agit de mots à résidus positifs, plus polysémiques qu'on n'aurait cru.

- *Mots à résidus positifs*

Les 158 spécificités ou mots à résidus positifs importants se caractérisent par un résidu (ou erreur d'estimation) supérieur à 1950. Ces mots sont peu spécifiques (rangs de spécificité > 3500) et plutôt hétérogènes sémantiquement (rangs de monosémie > 3500). En plus, ils ont une fréquence très élevée dans le corpus technique (rangs de fréquence technique < 1000), ainsi que dans le corpus général (rangs de fréquence générale < 1000 et souvent même < 500). Ce sont donc les spécificités les plus fréquentes du corpus technique et du corpus général[168].

Leur position dans la partie supérieure droite de la représentation s'explique donc par leur fréquence. Plus leur fréquence technique est importante, plus les mots auront des cooccurrents et des cooccurrents des cooccurrents pertinents et plus ils auront tendance à être hétérogènes sémantiquement, ce qui les situe dans la partie supérieure de la visualisation. En plus, ces mots se situent à droite en raison de leur fréquence générale très élevée, celle-ci étant responsable de leur degré de spécificité limité. En effet, le degré de spécificité est le rapport entre la fréquence technique relative et la fréquence générale relative. Si un mot est relativement plus fréquent dans le corpus technique, il sera spécifique du corpus technique et son degré de spécificité sera élevé (un rang de spécificité près de 1). Par contre, si un mot est relativement aussi fréquent ou un peu plus fréquent dans le corpus technique que dans le corpus général[169], il s'agit d'un cas limite des spécificités. C'est un mot plutôt général, fréquent dans le corpus général, mais quand même légèrement spécifique du corpus technique, car il est statistiquement significatif, bien qu'il frôle le seuil de significativité des spécificités. Un mot relativement moins fréquent dans le corpus technique que dans le corpus général n'est pas une spécificité (ou mot-clé) du corpus technique et dès lors, de tels mots ne figurent pas dans la liste de spécificités.

Il s'ensuit que les mots à résidus positifs importants, tels que *service, objet, commercial, air, bénéficier, intervenir*, sont des mots généraux, qui appartiennent à la langue générale, mais qui sont quand même plutôt fréquents dans le corpus technique (Cf. tableau 7.6).

---

[168] Il est à noter que ces 158 mots à résidus positifs importants se caractérisent par une très bonne corrélation entre le rang de fréquence technique et le rang de fréquence générale, à savoir un coefficient de corrélation de 0,98. Par contre, l'ensemble des 4717 spécificités se caractérise par un coefficient de corrélation de 0,76.

[169] Rapport *freqrel1/freqrel2* près de 1 (*freqrel1* = fréquence technique relative et *freqrel2* = fréquence générale relative).

rang_v_ spec	spécificité	résidus	rang_v_ freq1	rang_v_ freq2	rang_v_ mono	freqrel1/ freqrel2	nombre de c à 0,9999
4570	*service*	3972,32	143	13	4692	1,0825	115
4413	*objet*	3715,33	417	101	4550	1,1554	76
4470	*air*	3704,08	357	73	4497	1,1335	105
4278	*commercial*	3614,46	686	208	4548	1,2475	59
4347	*certain*	3473,99	71	7	4357	1,0817	71
4651	*fût*	3466,64	2548	1280	4127	1,7862	27
4634	*bénéficier*	3358,19	482	120	4031	1,1329	29
4605	*informer*	3341,95	1142	464	4036	1,2775	51
4580	*obligation*	3297,64	739	228	4010	1,1920	48
4588	*agir*	3279,50	234	26	3986	1,0958	38
4604	*récent*	3273,22	668	191	3968	1,1729	32
4432	*intervenir*	3205,25	592	160	4026	1,1820	35
3878	*demande*	3185,50	415	104	4412	1,1861	51
3666	*usine*	3110,24	713	243	4492	1,3171	47
3630	*provoquer*	3100,87	493	133	4509	1,2497	65
3729	*salon*	3090,38	920	361	4426	1,3749	50
4711	*correspondant*	3083,58	576	146	3700	1,1370	43
3680	*non*	3082,49	88	9	4454	1,1111	130
4666	*ouvert*	3074,63	945	333	3724	1,2110	22
4326	*clé*	3067,61	920	340	3966	1,3017	27
4469	*pro*	3057,35	1523	681	3851	1,4367	31

Tableau 7.6 Mots à résidus positifs les plus importants (supérieurs à 3000)

- *Mots à résidus négatifs*

Les 47 spécificités ou mots à résidus négatifs importants ont un résidu (erreur d'estimation) inférieur à -1950. Ces mots se caractérisent par une spécificité moyenne (rangs de spécificité entre 1000 et 2500) et par une homogénéité sémantique importante (rangs de monosémie < 1100) par rapport à leur rang de spécificité et par rapport à leur rang de fréquence technique (entre 1600 et 3500). Leur position dans la partie inférieure gauche de la représentation s'explique également par leurs caractéristiques linguistiques. Ces mots se trouvent à gauche, parce qu'ils sont plutôt spécifiques du corpus technique, ce qui s'explique par leur absence du corpus général. Ils se situent en bas, en raison de leur nombre limité de c (et de cc) significatifs, donc ils sont peu hétérogènes sémantiquement, en dépit de leur fréquence technique considérable.

Par conséquent, les mots à résidus négatifs importants sont des mots techniques, absents du corpus de langue générale (Cf. tableau 7.7).

rang_v _spec	spécificité	résidus	rang_v _freq1	rang_v _freq2	rang_v_ mono	nombre de c à 0,9999
1543	*électrobroches*	-2719,62	2429	3197	217	2
1885	*autocalibrage*	-2667,15	2796	3197	19	4
1534	*compacité*	-2656,22	2248	2618	287	2
2130	*équerrage*	-2466,71	3010	3197	40	4
2221	*reconditionnement*	-2423,06	3089	3197	17	5
2130	*nervurage*	-2421,71	3010	3197	85	2
1587	*hydrauliquement*	-2403,40	2489	3197	501	3
2041	*porte-fraise*	-2366,89	2920	3197	205	2
1928	*polygonal*	-2357,65	2725	2971	297	2
1740	*ablocage*	-2332,34	2651	3197	460	3
1986	*balayage*	-2329,18	2429	2306	283	4
1956	*goulotte*	-2311,15	2854	3197	323	2
2130	*détalonnage*	-2293,71	3010	3197	213	2
2041	*dynamiquement*	-2261,89	2920	3197	310	5
2467	*semi-conducteur*	-2257,90	3318	3197	2	2
2467	*servos*	-2257,90	3318	3197	2	3
2322	*annulaire*	-2242,09	2854	2618	124	2
1619	*rationnel*	-2229,96	1750	1544	651	3
2078	*équipé*	-2221,80	2137	1763	323	2
2389	*crique*	-2219,02	2599	2144	98	3
2345	*act/sign*	-2215,25	3199	3197	134	4

Tableau 7.7 Mots à résidus négatifs les plus importants (inférieurs à -2200)

Le tableau synoptique ci-dessus (Cf. tableau 7.8) permet de comparer les caractéristiques linguistiques des deux groupes de mots à résidus importants (positifs et négatifs) avec celles des 4717 spécificités.

	158 mots à résidus positifs	47 mots à résidus négatifs	4717 spécificités
rang_v_spec	> 2000 (PAS spécif)	1000-2500	0-4717
rang_v_mono	> 3000 (poly)	< 1100 (très mono)	0-4717
rang_v_freq1	< 1000 (fréq. tech.)	1600-3500	0-4717
rang_v_freq2	<500 (fréq. gén. !!!)	1400-3197 (bcp 3197)	0-4717
LLR	20-4 (très bas)	86-22	50000-4
résidus	>1950	< -1950	-2719 à 3972
nombre de c	> 25 (souvent)	2-6	2-390
corr rang_v_freq1~2	0,98	0,73	0,76

Tableau 7.8 Comparaison des mots à résidus importants et des 4717 spécificités

7.1.3.2 Exploration en fonction de la fréquence technique et générale

L'analyse exploratoire des mots à résidus importants, responsables du problème de l'hétéroscédasticité, a démontré le rôle de leur fréquence technique et générale. Nous proposons dès lors de procéder à une deuxième exploration en fonction de la fréquence technique et générale des 4717 spécificités. A cet effet, les 4717 spécificités sont réparties en 4 groupes en fonction de leurs rangs de fréquence technique et de fréquence générale, dans le corpus technique et dans le corpus général (Cf. tableau 7.9).

Il est clair que les mots à résidus positifs importants, les mots généraux, se trouvent dans le groupe 3. Les mots à résidus négatifs importants, les mots techniques, se situent dans le groupe 2. Les détails de ces 4 groupes de fréquence sont expliqués dans le document en annexe (Cf. annexe 11).

	groupe 1	groupe 2	groupe 3	groupe 4	référence
nombre de mots	1647	556	365	2149	4717
fréquence technique	+	**+**	**-**	-	
fréquence générale	+	**-**	**+**	-	
rang_v_freq1	1-2174	1-2174	2204-4284	2204-4284	1-4284
rang_v_freq2	1-2000	**2013-3179**	**1-2000**	2013-3179	1-3179
rang_v_spec	1-4717	3-1787	1952-4714	1347-4712	1-4717
	très fq	**mots tech. spécifiques**	**mots gén. peu spéc.**	peu fq peu spéc.	

Tableau 7.9 Répartition des 4717 spécificités en 4 groupes

7.1.4 Solutions et interprétations

Les explorations ayant conduit à mieux comprendre et à caractériser les spécificités responsables de l'hétéroscédasticité, nous procédons dans cette section aux solutions techniques habituelles pour traiter le problème de l'hétéroscédasticité et aux solutions alternatives (Cf. 7.1.5 et 7.1.6).

7.1.4.1 Solutions techniques

Les solutions techniques généralement adoptées consistent soit en des transformations logarithmiques ou polynomiales, soit en une analyse de régression simple pondérée, soit en une analyse de régression non linéaire. Nous commenterons ci-dessous les principaux résultats de ces trois solutions techniques, ainsi que leurs avantages et inconvénients. On pourrait éventuellement aussi envisager une analyse de régression logistique, normalement pour une variable dépendante binaire (0 / 1).

- *Transformations logarithmiques et polynomiales*

Les transformations logarithmiques d'une variable ou des deux variables permettent de résoudre le problème de l'hétéroscédasticité si les résidus se caractérisent par un patron sous forme d'entonnoir, donc par une augmentation progressive des résidus tant positifs que négatifs. Cela signifie que le rapport entre les variables n'est pas linéaire, mais logarithmique. La visualisation des résidus (Cf. figure 7.3) et la représentation simplifiée des résidus (Cf. figure 7.5) montrent un patron de dilatation (ou d'entonnoir), mais uniquement à droite de la visualisation.

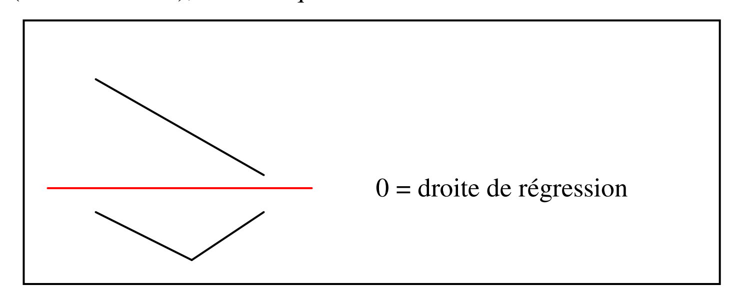


Figure 7.5 Représentation simplifiée des résidus

Pour les mots à droite de cette représentation simplifiée des résidus (les mots les plus spécifiques, tels que *m/mn, désiré, cavité, électrobroches*), une transformation logarithmique serait envisageable. Toutefois, les mots à gauche de cette représentation (les mots les moins spécifiques, comme *service, commercial*) ne s'y prêtent pas, parce que le patron d'entonnoir ne se prolonge pas. Le coin inférieur gauche de la représentation est vide parce que les mots à la frontière se situent à la limite de la significativité de la spécificité. Par conséquent, les transformations logarithmiques des variables ne permettent pas de résoudre le problème de l'hétéroscédasticité.

Les transformations polynomiales d'une variable ou des deux variables consistent à les élever au carré ou à la puissance *n* ou à extraire la racine carrée ou $n^{\text{ième}}$. Ces transformations conviennent surtout à des variables qui ne présentent pas de rapport linéaire, mais un rapport exponentiel, par exemple. La double transformation polynomiale ($y^2 \sim \sqrt{x}$) aboutit à l'homoscédasticité des résidus et à un pourcentage de variation expliquée R^2 de 57,38% (supérieur au pourcentage de 51,57% de la régression linéaire simple sans transformations). Du point de vue technique, le problème de l'hétéroscédasticité est résolu, mais du point de vue linguistique, il est plutôt difficile d'interpréter le carré du rang de monosémie et la racine carrée du rang de spécificité. En plus, la transformation polynomiale de la variable dépendante (rang de monosémie) est dangereuse, parce qu'elle peut avoir un impact sur les autres variables indépendantes, étant donné que leur rapport respectif avec la variable dépendante ne sera plus linéaire.

- *Analyse de régression pondérée*

La deuxième solution technique de la régression pondérée est généralement adoptée lorsque les résidus ne suivent pas de distribution normale. L'analyse de régression pondérée est basée sur la méthode des moindres carrés pondérés et consiste à accorder moins d'importance aux mots à résidus importants et plus d'importance aux mots à résidus limités. Le résultat de la régression pondérée[170] est un pourcentage de variation expliquée R^2 de 62,51%. En effet, la figure ci-dessous (Cf. figure 7.6) montre la tendance linéaire négative de manière nettement plus claire. L'analyse de régression pondérée confirme donc notre hypothèse initiale : les mots les plus spécifiques du corpus technique ne sont pas les plus monosémiques.

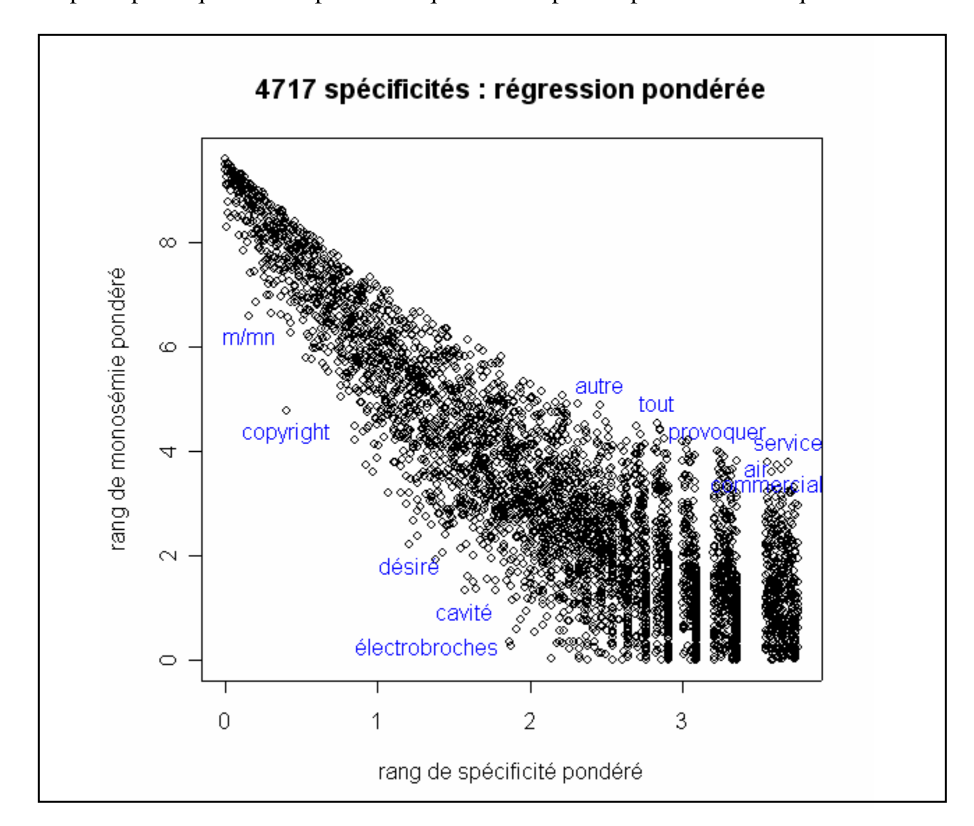


Figure 7.6 Régression pondérée : visualisation des résultats

[170] Fonction dans le logiciel R : `lm(y~x, weights=1/fitted(lm((resid(res)^2)~x))` avec `res=lm(y~x)`.

Comme le montre la figure 7.6, les mots sont nettement moins dispersés (d'où le R^2 plus élevé) et ils sont plus comprimés. Le mot *service*, par exemple, se trouve au même rang de spécificité, mais à un rang de monosémie estimé plus bas, donc plus monosémique. Les mots à résidus positifs importants sont des mots généraux (peu spécifiques) et polysémiques. Une pondération en fonction de leurs résidus importants les ramène plus près de la droite de régression, à un rang moins polysémique, suivant la tendance générale de rapport linéaire négatif. Bien entendu, cela revient à sous-estimer ou à méconnaître la fréquence technique importante de ces mots et le nombre important de cooccurrents, qui leur confèrent leur statut polysémique. En dépit cette sous-estimation des caractéristiques de certains mots, l'analyse de régression pondérée confirme notre hypothèse initiale.

- *LOESS ou l'analyse de régression non linéaire*

Lorsque le rapport entre les deux variables n'est pas vraiment linéaire, la technique LOESS de régression non linéaire permet de visualiser le rapport (non linéaire) entre les deux variables. Cette technique est purement visuelle et n'aboutit donc pas à un pourcentage de variation expliquée. Le résultat de LOESS[171] ou des régressions locales n'est pas une droite (linéaire), mais une courbe, visualisée par la figure ci-dessous (Cf. figure 7.7). Pour les mots les plus spécifiques, à gauche de la visualisation, cette courbe ressemble beaucoup à la droite de régression de l'analyse de régression linéaire simple (Cf. figure 7.1). En effet, au début, pour les mots les plus spécifiques du corpus technique, la courbe s'incline vers le bas et visualise clairement le rapport négatif entre le rang de spécificité et le rang de monosémie. Toutefois, pour les mots moins spécifiques, c'est-à-dire à partir du rang de spécificité 3000 et plus clairement encore à partir de 3500, la tendance négative de la courbe descendante s'estompe. Elle tend même à s'inverser en une tendance légèrement positive pour les rangs de spécificité supérieurs à 4000, tenant compte ainsi des mots peu spécifiques plutôt polysémiques (à résidus importants) qui se situaient loin de la droite de régression de la figure 7.1.

[171] LOESS (*Local Polynomial Regression Fitting*) permet d'ajuster (*fit*) une courbe à travers un nuage de points, à partir d'un ajustement local (*local fitting*). Pour un ajustement au point *x*, LOESS utilise des points dans le voisinage de *x*, pondérés en fonction de leur distance de *x*.

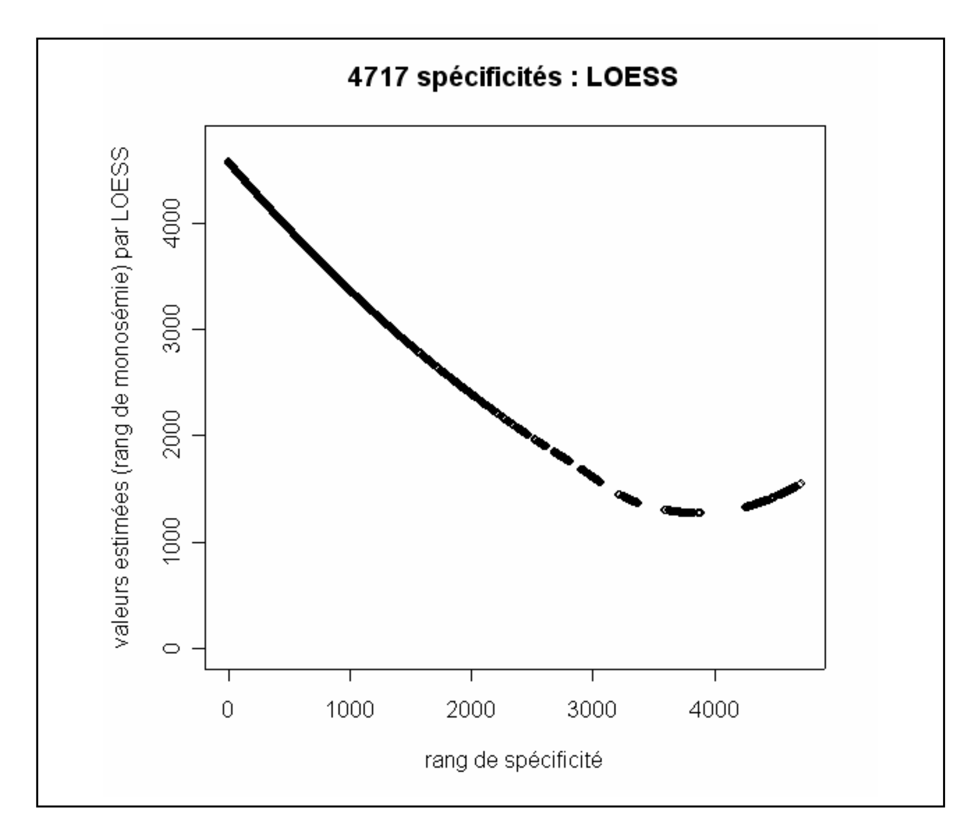


Figure 7.7 Régression non linéaire : visualisation de LOESS

En effet, pour les mots les moins spécifiques (et les plus généraux) à droite de la visualisation, la courbe remonte et visualise un faible rapport positif, ce qui signale un léger effet de la thèse monosémiste. Il est clair que les mots les moins spécifiques (et les plus généraux) échappent à la tendance générale du rapport négatif entre le rang de monosémie et le rang de spécificité. Par conséquent, ils échappent à la corrélation linéaire négative et à la capacité de prédiction du modèle de régression linéaire général.

Les mots peu spécifiques (à droite de la visualisation) sont généralement assez monosémiques et se situent en bas de la visualisation. Néanmoins, une fraction de ces mots peu spécifiques du corpus technique consiste en des mots de la langue générale. Ces mots peu spécifiques et plus généraux se caractérisent par la polysémie, confirmant ainsi la thèse des monosémistes (polysémie des mots de la langue générale). Toutefois, ces mots ne représentent qu'une fraction limitée des 4717 spécificités et ne remettent pas en question la conclusion générale du rapport négatif entre le rang de monosémie et le rang de spécificité.

7.1.4.2 Solution de répartition en plusieurs sous-ensembles

Les solutions techniques commentées ci-dessus apportent certes une solution technique au problème de l'hétéroscédasticité et permettent d'aboutir à des pourcentages de variation expliquée plus élevés. Elles semblent en outre confirmer notre hypothèse. Toutefois, ces solutions sont discutables du point de vue linguistique, parce qu'il n'est pas clair comment on pourrait interpréter le carré du rang de monosémie et la racine carrée du rang de spécificité. La régression pondérée reviendrait à sous-estimer l'importance des caractéristiques linguistiques (de fréquence) de certaines spécificités. La régression non linéaire LOESS, quant à elle, visualise le rapport non linéaire, confirme la tendance négative pour les mots spécifiques et indique que celle-ci ne s'applique toutefois pas à toutes les spécificités. En effet, les mots généraux et peu spécifiques échappent à la règle.

Les solutions techniques ont donc permis d'avancer dans l'analyse dans la mesure où elles montrent que la tendance linéaire négative ne convient pas à toutes les 4717 spécificités, mais peut-être à un sous-ensemble. Cette observation nous conduit à répartir les spécificités en plusieurs sous-ensembles, en fonction de différents critères de répartition. Ces répartitions visent donc principalement à vérifier le rapport entre le rang de monosémie et le rang de spécificité par sous-ensemble, en termes de variation expliquée (R^2), de type de corrélation (négative ou positive) et d'homoscédasticité.

Les explorations préliminaires et les solutions techniques ont mis en évidence l'importance de la fréquence technique et de la fréquence générale des spécificités, qui constituent donc des critères de répartition importants. En plus, nous aimerions exploiter l'écart entre le rang de fréquence technique d'un mot spécifique et son rang de fréquence générale. Cet écart constitue un critère de technicité (mots plus ou moins techniques) et conduit à l'élaboration d'une nouvelle variable, à savoir l'écart des rangs de fréquence. Celui-ci permettra également d'effectuer des répartitions supplémentaires.

- *Une nouvelle variable : l'écart des rangs de fréquence*

Afin de déterminer la valeur de la nouvelle variable, nous déterminons la différence ou l'écart entre le rang de fréquence technique et le rang de fréquence générale[172] des 4717 spécificités.

Les rangs de fréquence technique se situent entre 1 et 4284, étant donné que les spécificités avec une fréquence absolue identique se verront attribuer un rang de fréquence identique. Un rang près de 1 signifie que la spécificité en question est très fréquente dans le corpus technique. Comme les hapax ont été supprimés de la liste des spécificités, 434 spécificités ont la fréquence technique minimale de 2. Les rangs de fréquence générale se situent entre 1 et 3197, des rangs identiques étant attribués à des spécificités avec la même fréquence absolue. Un nombre très important de spécificités (1521) ne figurent pas dans le corpus de langue générale et se caractérisent donc par une fréquence absolue dans le corpus général de zéro. Par conséquent, il y a moins de rangs différents de fréquence générale (3197) que de rangs de fréquence technique (4284).

La solution consiste à rééchelonner ou à réencoder la variable du rang de fréquence générale (rang_v_freq2), donc à multiplier les valeurs par 4284 et à diviser par 3197, c'est-à-dire à recourir au facteur 1,34. La nouvelle variable, l'écart des rangs de fréquence, correspond à la différence entre le rang de fréquence générale rééchelonné et le rang de fréquence technique. Les valeurs numériques de la nouvelle variable (ecart_r_v_freq) sont soit positives si les spécificités sont beaucoup plus fréquentes dans le corpus technique (rang_v_freq1 très bas et près de 1), soit négatives, si les spécificités sont des mots de la langue générale (rang_v_freq2 très bas et près de 1). Un écart des rangs de fréquence autour de zéro signifie que la spécificité se caractérise par des rangs de fréquence technique et générale comparables (compte tenu du rééchelonnement du rang de fréquence générale).

---

[172] Il s'agit des variables *rang_v_freq1* et *rang_v_freq2* (rang de fréquence identique pour des valeurs de fréquence absolue identiques). Par contre, *rang_freq1* et *rang_freq2* indiquent le classement des spécificités, de 1 à 4717, en fonction de leur fréquence absolue dans le corpus technique (*freqabs1*) et en fonction de leur fréquence absolue dans le corpus général (*freqabs2*) (sans tenir compte de valeurs identiques, donc sans rangs identiques).

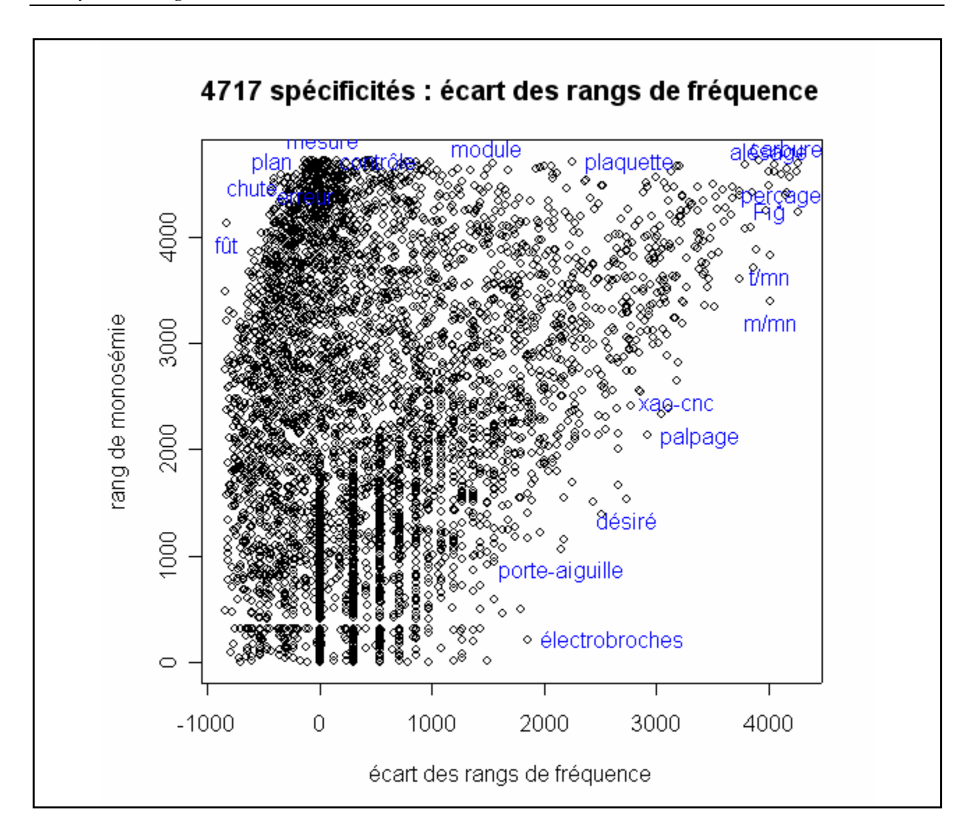


Figure 7.8 Visualisation de l'écart des rangs de fréquence

La visualisation ci-dessus (Cf. figure 7.8) montre clairement que les mots plus techniques se situent plus à droite, caractérisés par un écart positif et que les mots plus généraux se situent à gauche, avec un écart négatif. Par conséquent, l'écart des rangs de fréquences constitue un critère supplémentaire pour caractériser les 4717 spécificités, techniques ou générales.

En effet, la figure 7.9 ci-dessous montre les deux critères permettant de caractériser les spécificités, à savoir le degré de spécificité visualisé par log(LLR)[173] et l'écart des rangs de fréquence.

[173] Le log du degré de spécificité (log(LLR)) permet de rééchelonner les degrés de spécificité ou les valeurs de LLR, qui s'étendent entre 50521 (log = 4,70) et 3,85 (log = 0,58).

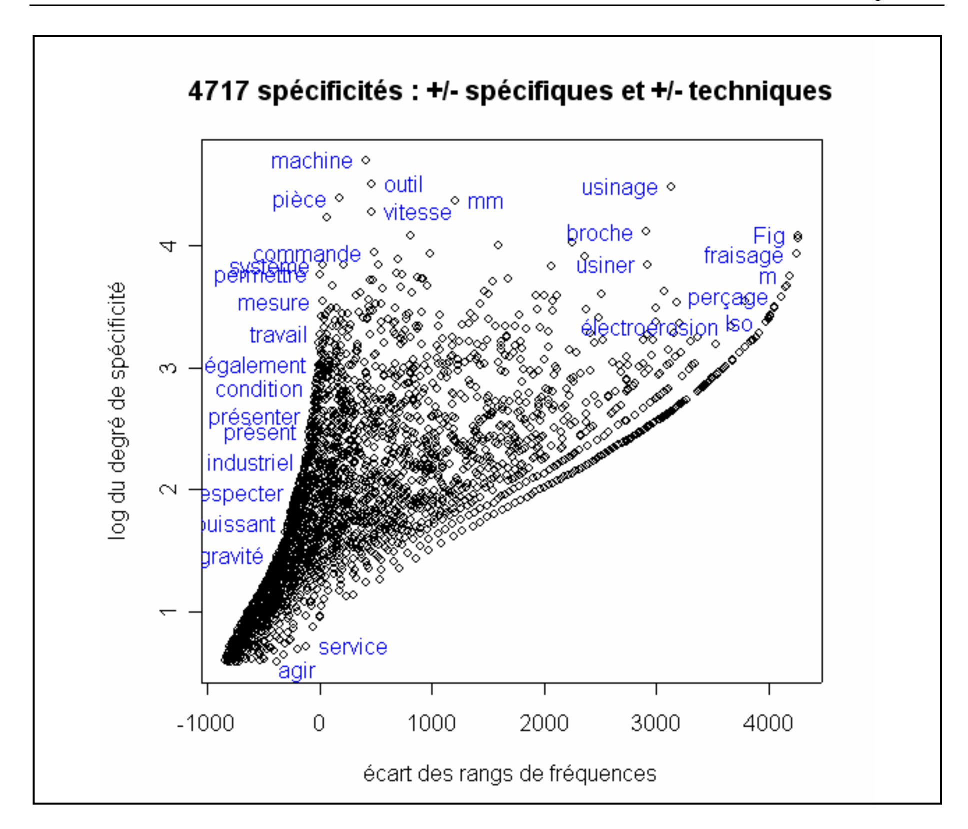


Figure 7.9 Spécificités plus et moins spécifiques et techniques

Les mots les plus spécifiques du corpus technique se trouvent en haut de la visualisation, les mots les moins spécifiques se trouvent en bas, en fonction de l'axe vertical de la spécificité. L'axe horizontal de l'écart des rangs de fréquence permet de faire des distinctions supplémentaires pour des mots avec un degré de spécificité comparable, telles que la distinction intéressante entre *fraisage* (plus technique) et *permettre* (plus général). La prise en compte de ces deux critères, le degré de spécificité et l'écart des rangs de fréquence[174], conduit donc à une granularité plus fine des caractéristiques des spécificités.

[174] Il est à noter que le degré de spécificité est calculé à partir de la significativité statistique de la différence des fréquences relatives technique et générale. L'écart des rangs de fréquence est calculé à partir de la différence entre les deux rangs de fréquence (après rééchelonnement du rang de fréquence générale).

- *Répartition en fonction du rang de fréquence technique*

Dans un premier temps, les 4717 spécificités du corpus technique seront réparties en trois groupes ou sous-ensembles en fonction de leur rang de fréquence technique et en fonction des deux critères établis ci-dessus (log du LLR et écart des rangs de fréquence).

A cet effet, le rang de fréquence technique est visualisé comme troisième variable sur la visualisation précédente (Cf. figure 7.9), au moyen de 4284 couleurs[175] (Cf. annexe 12 : figure A12.1). Les bandes de couleur du rang de fréquence technique sont orientées plutôt horizontalement et suivent donc les degrés de spécificité (log du LLR). On fait la distinction entre 3 groupes de fréquence technique, à savoir les spécificités les plus fréquentes du corpus technique (rangs de fréquence technique entre 1 et 782), les spécificités moyennement fréquentes (rangs entre 786 et 2321) et les spécificités les moins fréquentes du corpus technique (rangs de 2368 à 4284).

	rangs	spécificités	R²	homoscédasticité ?
rvfq1_A	1-782	785	5,45%	hétéroscédasticité
rvfq1_B	786-2321	1582	8,13%	hétéroscédasticité
rvfq1_C	2368-4284	2350	17,23%	homoscédasticité

Tableau 7.10 Spécificités : 3 groupes de rang de fréquence technique

Le tableau ci-dessus (Cf. tableau 7.10) indique, par groupe ou sous-ensemble de rang de fréquence technique, le nombre de spécificités, le pourcentage de variation expliquée R^2 (de la régression linéaire simple entre le rang de monosémie et le rang de spécificité pour le sous-ensemble) et finalement l'homoscédasticité éventuelle. La corrélation entre le rang de monosémie et le rang de spécificité par sous-ensemble est toujours négative.

Cette répartition en fonction du rang de fréquence technique ne s'avère pas satisfaisante. D'une part, les pourcentages de variation expliquée R^2 sont trop faibles. D'autre part, dans deux groupes sur trois, l'hétéroscédasticité des résidus pose problème. En plus, les trois sous-ensembles suivent les axes X et Y des rangs de monosémie et de spécificité de la visualisation de la régression linéaire de base. Le groupe A est, en gros, le plus spécifique et le plus polysémique, le groupe B se situe au milieu pour les deux variables, alors que le groupe C est le moins spécifique et le plus monosémique. Une autre répartition des spécificités s'impose donc, par exemple en fonction du rang de fréquence générale.

---

[175] Il y a autant de couleurs que de rangs de fréquence technique (gamme de couleurs : arc-en-ciel) : `col=rainbow(4284)`.

- *Répartition en fonction du rang de fréquence générale*

Pour le rang de fréquence générale, nous procédons de la même façon. Le tableau ci-dessous (Cf. tableau 7.11) montre les résultats pour les trois sous-ensembles de spécificités en fonction des bandes de fréquence générale (Cf. annexe 12 : figure A12.2). Généralement, les pourcentages de R^2 sont plus élevés que ceux des sous-ensembles en fonction du rang de fréquence technique. Notons surtout le pourcentage de R^2 élevé (63,23%) des mots les moins fréquents ou même absents du corpus de langue générale (rvfq2_C). Les spécificités les plus fréquentes dans le corpus général, par contre, n'ont pas de bonne corrélation entre le rang de monosémie et le rang de spécificité et se caractérisent d'ailleurs par l'hétéroscédasticité. Les observations de cette répartition confirment les résultats de l'analyse de régression linéaire simple de base et des solutions techniques commentées ci-dessus (Cf. 7.1.4.1).

	rangs	spécificités	R^2	homoscédasticité ?
rvfq2_A	1-784	785	33,81%	hétéroscédasticité
rvfq2_B	786-1871	1099	45,34%	homoscédasticité
rvfq2_C	1885-3197	2833	63,23%	homoscédasticité

Tableau 7.11 Spécificités : 3 groupes de rang de fréquence générale

Comme les groupes A et B ne comprennent pas beaucoup de spécificités et que A est sujet à l'hétéroscédasticité, nous avons décidé de les regrouper. Le sous-ensemble rvfq2_AB comprend 1884 spécificités, affiche un R^2 de 40,37%, mais est toujours sujet à de l'hétéroscédasticité, qui semble hanter les mots les plus généraux parmi les spécificités peu spécifiques.

En revanche, la répartition des spécificités en trois sous-ensembles équilibrés en fonction du rang de fréquence générale aboutit à de meilleurs résultats, plus équilibrés, visualisés par le tableau 7.12 ci-dessous. Cependant, le problème de l'hétéroscédasticité se pose toujours pour le sous-ensemble (rvfq2_a) des spécificités les plus générales, c'est-à-dire les plus fréquentes dans le corpus général.

	rangs	spécificités	R^2	homoscédasticité ?
rvfq2_a	1-1555	1564	38,84%	hétéroscédasticité
rvfq2_b	1565-2800	1406	60,54%	homoscédasticité
rvfq2_c	2971-3197	1747	67,71%	homoscédasticité

Tableau 7.12 Spécificités : 3 groupes équilibrés de rang de fréquence générale

- *Répartition en fonction des coupes de spécificité et de technicité*

Afin de répartir les spécificités en fonction des critères de spécificité (log du LLR) et de technicité (écart des rangs de fréquence), nous proposons de procéder à des coupes, visualisées par la figure ci-dessous (Cf. figure 7.10). A partir des coordonnées de deux points et des valeurs Y à l'origine (x = 0), les droites des deux coupes sont identifiées[176] et permet la répartition des spécificités en trois sous-ensembles.

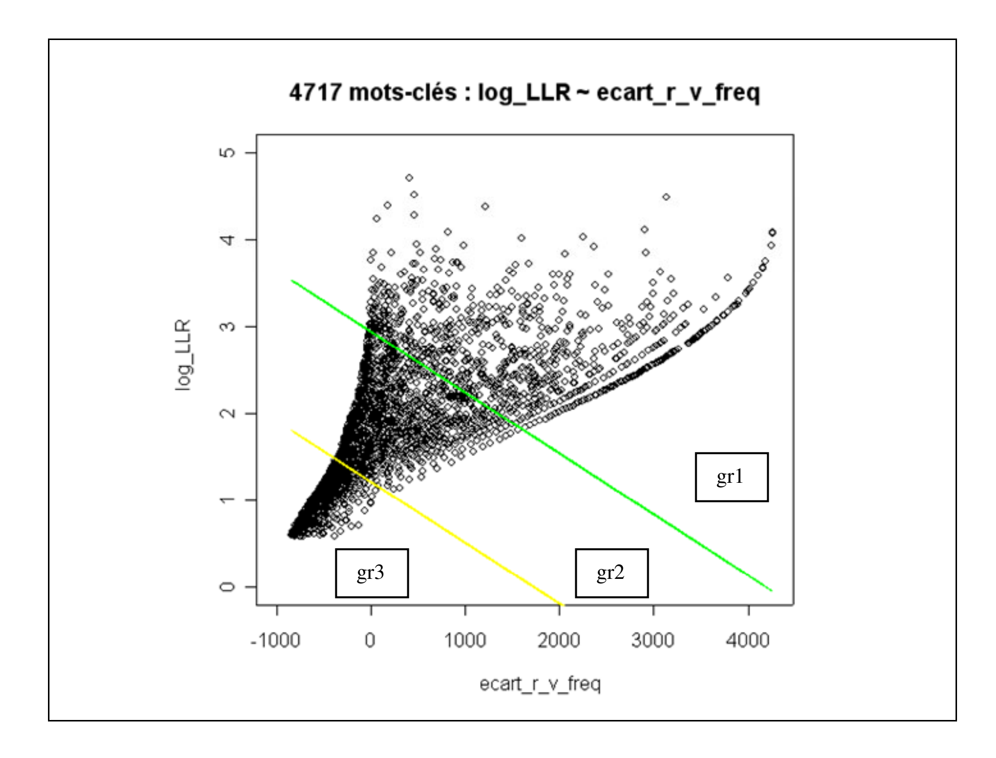


Figure 7.10 Visualisation des coupes : spécificité et technicité

Le groupe 1 se situe en haut à droite et comprend les mots les plus spécifiques et les plus techniques, le groupe 2 se situe au milieu et le groupe 3 se trouvant en bas à gauche comprend les mots les moins spécifiques et les plus généraux (les moins techniques). Le tableau ci-dessous (Cf. tableau 7.13) visualise les résultats pour les trois groupes répartis par les coupes en fonction des deux critères de spécificité et de technicité.

---

[176] Les droites des deux coupes sont perpendiculaires à la médiane à travers le nuage de points.

	rangs	spécificités	R^2	homoscédasticité ?
groupe1	+ spec + tech	1063	68,08%	homoscédasticité
groupe2	+/-spec +/-tech	2295	59,9%	homoscédasticité
groupe3	- spec - tech	1359	10,39%	homoscédasticité

Tableau 7.13 Spécificités : 3 groupes de spécificité et de technicité

Ces résultats montrent que le groupe 3 des mots peu spécifiques et très généraux se caractérise par un pourcentage très faible de variation expliquée R^2 (10%). Les deux autres groupes affichent de bons résultats. En plus, la répartition en fonction de ces deux critères permet de résoudre le problème de l'hétéroscédasticité. Si les spécificités sont triées par ordre ascendant de coupe, donc en fonction du critère de répartition, le gqtest confirme l'homoscédasticité dans les trois groupes. Le problème de l'hétéroscédasticité étant résolu, cette répartition des spécificités aboutit à des résultats satisfaisants pour les deux premiers groupes, à savoir une corrélation linéaire négative entre le rang de monosémie et le rang de spécificité. Toutefois, le dernier groupe des spécificités les moins spécifiques et les plus générales ne se prête pas à une prédiction du rang de monosémie à partir du rang de spécificité.

- *Répartition en fonction de l'écart des rangs de fréquence*

Un dernier critère de répartition des spécificités, intéressant en fonction de l'interprétation des données, est celui de l'écart des rangs de fréquence. A cet effet, nous faisons la distinction entre les spécificités à écart positif (les spécificités techniques), les spécificités à écart négatif (les spécificités générales) et les spécificités qui se situent autour de zéro[177] (-50 et +50). Autour de zéro, différents seuils de *cut off* ont été testés et l'intervalle de -50 et +50 génère les meilleurs résultats. Un intervalle plus large (par exemple -100 et +100) revient à inclure des spécificités légèrement plus techniques et légèrement plus générales.

Le tableau ci-dessous (Cf. tableau 7.14) visualise les résultats pour les trois sous-ensembles, tant pour l'intervalle de 50 autour de zéro que pour l'intervalle de 100 autour de zéro. Les spécificités de chaque sous-ensemble sont classées par ordre décroissant d'écart des rangs de fréquence. Les résultats pour les deux seuils sont plutôt similaires. Dès lors, nous préférons maintenir le seuil le plus restreint (-50 et

[177] Si l'on considère séparément les spécificités ayant un écart de 0, l'analyse de régression est impossible pour ce groupe, étant donné que ces spécificités ont toutes le même rang de spécificité. Par conséquent, il faudra être moins sévère et inclure également des spécificités avec un écart positif et négatif très faible (autour de zéro). Ces spécificités constituent d'ailleurs un groupe très intéressant, étant donné qu'elles présentent tous les rangs de monosémie et tous les rangs de fréquence technique et générale.

+50 autour de zéro), qui intègre le moins de spécificités techniques et générales, à écart positif et négatif.

	écart	spécificités	R^2	homoscédasticité ?
ez50	autour de zéro	649	88,26%	homoscédasticité
ep50	positif > 50	2747	75,27%	homoscédasticité
en50	négatif < -50	1321	41%	homoscédasticité
ez100	autour de zéro	831	87,29%	homoscédasticité
ep100	positif > 100	2666	75,21%	homoscédasticité
en100	négatif < -100	1220	37,88%	homoscédasticité

Tableau 7.14 Spécificités : 3 groupes d'écart des rangs de fréquence

Il est clair que les spécificités autour de zéro (dont certaines sont légèrement techniques et d'autres plutôt générales) et les spécificités à écart positif (les plus techniques) se caractérisent par la meilleure corrélation négative entre le rang de monosémie et le rang de spécificité et donc par les R^2 les plus élevés, de 88,26% et 75,27% respectivement. De nouveau, les mots généraux (écarts négatifs à partir de -50) se caractérisent par un pourcentage de variation expliquée R^2 plutôt faible (41%). Par conséquent, la répartition des spécificités en trois sous-ensembles en fonction de l'écart des rangs de fréquence confirme la conclusion formulée ci-dessus pour les spécificités les plus générales : leur rang de spécificité ne rend compte qu'en partie de leur rang de monosémie.

L'écart des rangs de fréquence s'est avéré utile en tant que critère de répartition des données (spécificités techniques versus spécificités plutôt générales) et en tant que critère permettant des coupes en fonction des axes de spécificité et de technicité. Cependant, comme variable indépendante d'une analyse de régression simple, l'écart des rangs de fréquence est moins utile, étant donné sa corrélation très faible avec le rang de monosémie (coefficient de corrélation Pearson de 0,24).

Les solutions de répartition des données en plusieurs sous-ensembles, en fonction de différents critères de répartition, conduisent à quelques conclusions intéressantes. La corrélation linéaire négative entre le rang de monosémie et le rang de spécificité s'applique de manière inégale aux 4717 spécificités. En effet, le sous-ensemble des spécificités les plus générales (fréquence générale importante) et les moins spécifiques (du corpus technique) constitue une exception à la tendance générale de corrélation négative. Si ce sous-ensemble de spécificités peu spécifiques plutôt générales fait quand même l'objet d'une analyse de régression simple séparée, le pourcentage de variation expliquée R^2 est très faible, du point de vue de plusieurs critères de répartition. En plus, ce sous-ensemble est également sujet au problème de l'hétéroscédasticité des résidus, pour la plupart des critères de répartition. Par

conséquent, nous proposons d'élaborer une dernière solution qui consiste à exclure un sous-ensemble restreint de spécificités plutôt générales qui ne suivent pas la tendance générale et qui sont responsables de l'hétéroscédasticité de l'ensemble des 4717 spécificités.

7.1.4.3 Solution d'exclusion d'un sous-ensemble

Cette dernière solution vise à répartir les 4717 spécificités en deux sous-ensembles, dans le but, d'une part, d'exclure un sous-ensemble restreint de spécificités (plutôt générales) qui ne suivent pas la tendance générale et, d'autre part, de trouver un patron de base fiable pour les spécificités restantes, qui constituent le sous-ensemble plus étendu. Ce patron de base pourrait se présenter sous forme de corrélation négative, comme celle qui est visualisée par la droite de régression descendante. Pour le sous-ensemble restreint, un deuxième patron superposé pourrait se distinguer, différent du patron de base.

Des requêtes en Access à partir des valeurs des 4717 spécificités, permettent de recourir à plusieurs critères pour isoler un sous-ensemble restreint à exclure. Ainsi, un seuil de fréquence générale toujours plus bas, par exemple, permet d'identifier et d'exclure un sous-ensemble toujours plus grand de spécificités générales. Ensuite, pour le sous-ensemble de spécificités restantes, l'analyse de régression et le gqtest permettent de déterminer le pourcentage de R^2 et l'homoscédasticité. Ces opérations d'exclusion consécutives conduisent dès lors à établir une frontière nette entre l'hétéroscédasticité et l'homoscédasticité et à identifier avec précision et exactitude le sous-ensemble de spécificités responsable de l'hétéroscédasticité.

Différents critères se prêtent à ces expérimentations d'exclusion, à savoir la fréquence absolue dans le corpus technique, la fréquence absolue dans le corpus général, le degré de spécificité (valeur de LLR) et la diagonale (-1) des rangs de spécificité et des rangs de monosémie. Le critère potentiel de l'importance des résidus est rejeté, parce qu'il repose principalement sur les résultats de l'analyse de régression linéaire simple.

La fréquence absolue des spécificités dans le corpus technique ne permet pas d'isoler un sous-ensemble responsable de l'hétéroscédasticité. En effet, au fur et à mesure que le seuil de fréquence technique diminue, le sous-ensemble restant se caractérise par une fréquence technique toujours plus faible, mais aussi par un R^2 toujours plus bas. En plus, l'hétéroscédasticité ne disparaît pas.

En revanche, la fréquence absolue dans le corpus général conduit à des résultats nettement plus concluants et plus satisfaisants, parce que le seuil d'exclusion se situe à la fréquence absolue dans le corpus général de 52. C'est-à-dire que les 1507 spécificités dont la fréquence absolue dans le corpus général est supérieure ou égale

à 52 sont responsables de l'hétéroscédasticité. Le sous-ensemble restant des 3210 spécificités peu fréquentes ou même absentes du corpus général se caractérise par l'homoscédasticité et par un pourcentage de variation expliquée R² de 60,35% (Cf. figure 7.11). Les détails de cette expérimentation sont expliqués en annexe (Cf. annexe 12 : 12.4).

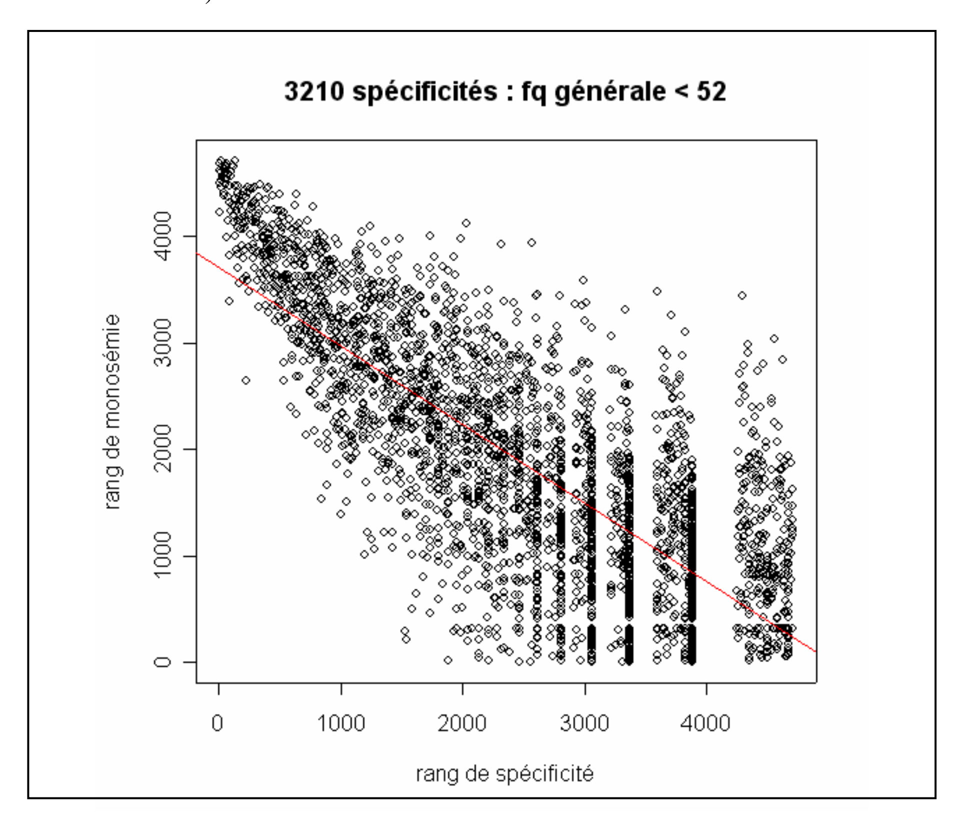


Figure 7.11 Exclusion d'un sous-ensemble : fréquence générale

Les autres critères évoqués ci-dessus ont également fait l'objet d'expérimentations d'exclusion. Si un seuil d'exclusion combiné de fréquence technique et générale, qui exclut les mots les plus fréquents dans les deux corpus, ne résout pas le problème, l'adoption du degré de spécificité comme seuil d'exclusion ne permet pas non plus de résoudre le problème. Comme les spécificités peu spécifiques sont exclues à partir de la droite vers la gauche, à partir des spécificités les moins spécifiques, l'hétéroscédasticité n'est pas résolue. Même si les mots en haut à droite sont exclus, c'est-à-dire les mots peu spécifiques et plutôt généraux, les mots en bas à droite sont exclus en même temps. Ceux-ci se situent aux alentours de la droite de régression et confirment la tendance générale de corrélation négative.

Finalement, la diagonale (droite à -1) des rangs de spécificité et de monosémie génère des résultats similaires à ceux de la fréquence absolue dans le corpus général (Cf. annexe 12 : 12.5), en ce qui concerne la taille et les résultats pour le sous-ensemble restant (R^2 et homoscédasticité). Toutefois, nous préférons adopter un critère d'exclusion indépendant des rangs de spécificité et de monosémie. Nous optons dès lors pour la fréquence générale (seuil de 52) pour l'exclusion du sous-ensemble.

7.1.5 Caractérisation du sous-ensemble exclu

Les expérimentations de répartition et d'exclusion ont démontré que les spécificités les moins spécifiques et/ou les plus fréquentes dans le corpus général se démarquent de la tendance générale de corrélation négative. En outre, l'exclusion du sous-ensemble restreint des 1507 spécificités les plus générales a permis d'aboutir à l'homoscédasticité du sous-ensemble des 3210 spécificités restantes. Dans cette section, les deux sous-ensembles feront l'objet d'une étude comparative, qui permet non seulement de relever leurs caractéristiques linguistiques respectives, mais aussi de fonder et de justifier l'interprétation linguistique du sous-ensemble des 3210 spécificités plutôt techniques (Cf. 7.1.6).

7.1.5.1 Caractéristiques principales des 1507 spécificités exclues

Le sous-ensemble des 1507 spécificités est responsable de l'hétéroscédasticité et de la perturbation de la tendance générale. Ce sont des mots fréquents dans le corpus général (fréquence absolue supérieure ou égale à 52), tant des mots peu spécifiques (à droite de la visualisation), tels que *service, objet* et *commercial*, que des mots très spécifiques (à gauche), comme *machine, outil* et *pièce*. (Cf. figure 7.12).

La figure 7.12 ci-dessous montre la droite de régression des 1507 spécificités (ligne continue) et, en tireté, la droite de régression des 4717 spécificités. Il est clair que les 1507 spécificités se situent majoritairement en haut de la droite de régression en tireté et qu'elles sont donc plus polysémiques, surtout les spécificités peu spécifiques (à droite). La droite de régression des 1507 spécificités (ligne continue) se situe aussi au-dessus de la droite de régression en tireté. Ces observations confirment la polysémie générale des 1507 spécificités, compte tenu de tous les cc (cc généraux et cc techniques), en dépit de la monosémie relative de quelques-unes de ces 1507 spécificités (en bas de la visualisation). On pourrait donc avancer l'hypothèse de la percolation de la polysémie générale des 1507 spécificités dans le corpus technique. Ce sont des mots généraux, polysémiques dans la langue générale, qui maintiennent leur polysémie lorsqu'ils sont employés dans un corpus technique.

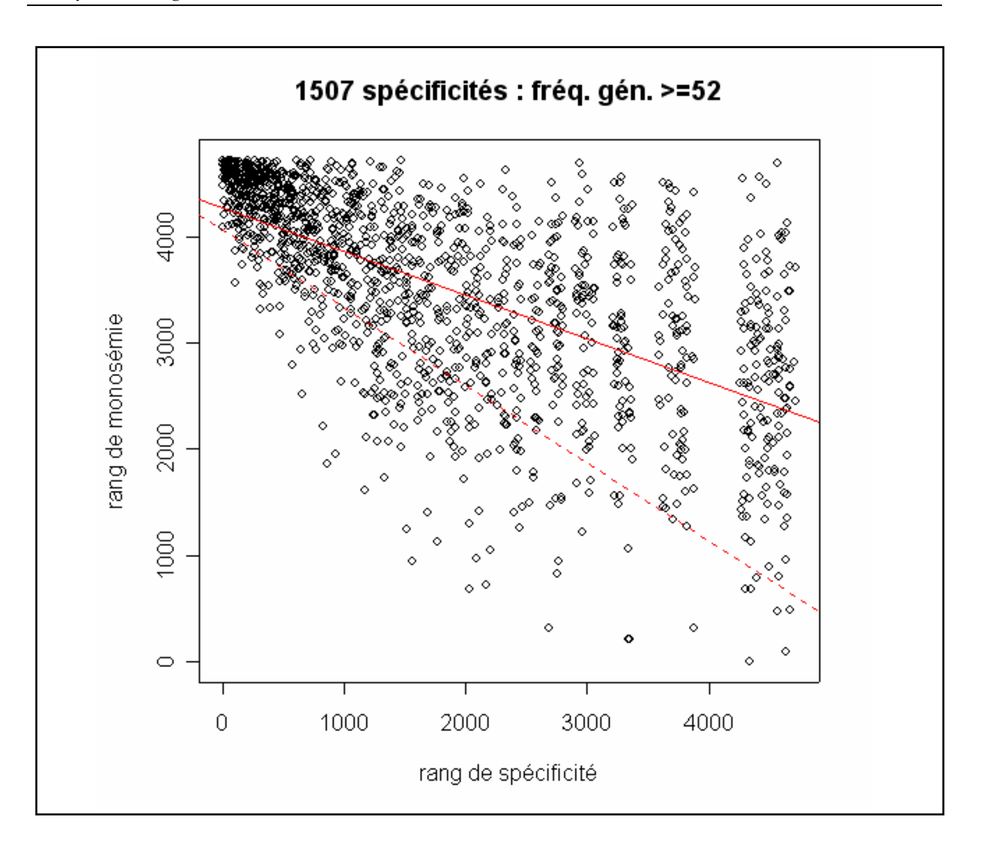


Figure 7.12 Sous-ensemble exclu (1507 spécificités) : monosémie

La visualisation ci-dessous des rangs de monosémie technique et des rangs de spécificité des 1507 spécificités (Cf. figure 7.13) confirme la tendance observée pour les rangs de monosémie : les spécificités se situent principalement en haut de la visualisation.

Toutefois, en ce qui concerne la différence entre la monosémie de base et la monosémie technique, les 1507 spécificités générales suivent assez bien la tendance des 4717 spécificités. En effet, si l'on tient compte de la technicité des cc, les spécificités les plus spécifiques (à gauche) sont un peu plus monosémiques (situées un peu plus en bas) par rapport à la visualisation des rangs de monosémie de base (Cf. figure 7.12) ; les spécificités les moins spécifiques (à droite) sont un peu plus polysémiques (situées un peu plus en haut). La figure 7.13 montre également, en pointillé, les droites de régression pour la monosémie de base (Cf. figure 7.12). Le déplacement des droites de régression pour la monosémie technique est similaire pour la ligne continue (1507) et pour la droite en tireté (4717). Il en ressort que les 1507 spécificités générales ne se caractérisent pas par une moindre polysémie technique, au contraire.

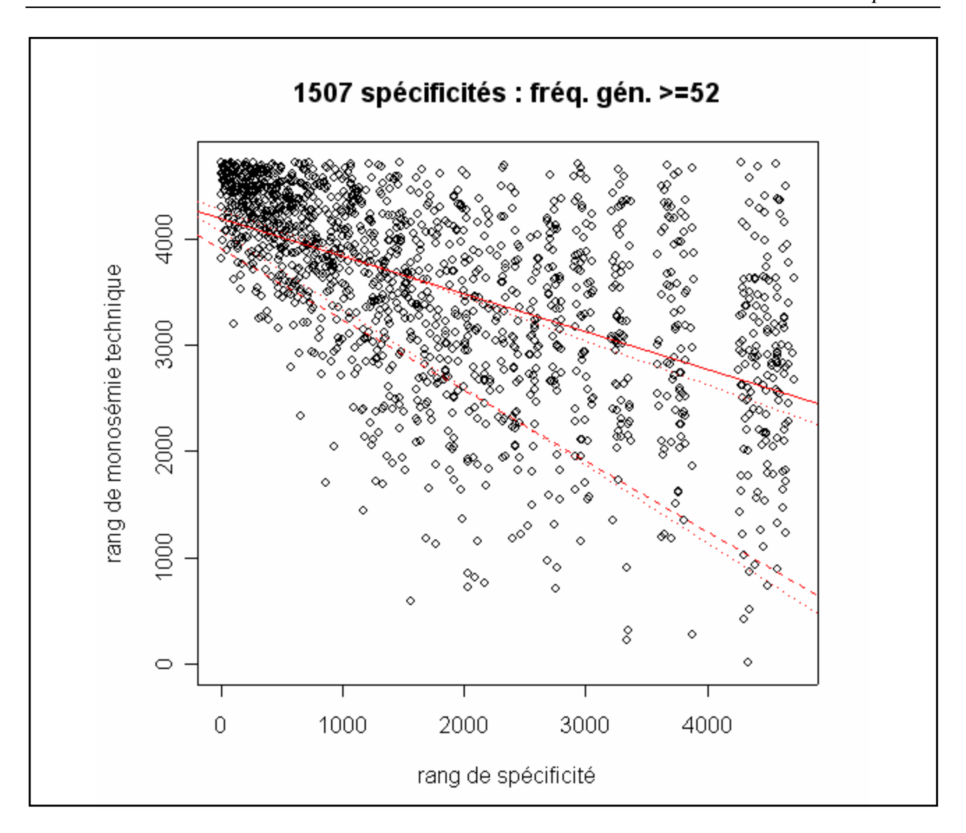


Figure 7.13 Sous-ensemble exclu (1507 spécificités) : monosémie technique

Si les 1507 spécificités générales se caractérisaient uniquement par une polysémie générale et pas ou très peu par une polysémie technique, leurs cc techniques se recouperaient. Par conséquent, la mesure de monosémie technique, en raison de sa pondération plus lourde pour les cc techniques, mènerait à des rangs de monosémie technique plus bas. Or, les 1507 spécificités ne se caractérisent pas par une monosémie technique plus importante. Elles manifestent, par contre, une polysémie technique plus importante, ceci étant surtout vrai pour les spécificités les moins spécifiques. Même si les spécificités les plus spécifiques (à gauche) deviennent légèrement plus monosémiques techniquement (leurs cc techniques se recoupent un peu), elles se situent toujours dans les rangs de monosémie technique plutôt élevés. En guise de conclusion, les 1507 spécificités se caractérisent par la polysémie générale (surtout pour les mots les plus spécifiques) et par la polysémie technique. L'analyse de l'impact combiné des facteurs caractérisant les cc permettra de vérifier et de nuancer cette conclusion (Cf. 7.1.5.2).

Avant de passer à l'analyse de l'impact combiné et aux corrélations de ces facteurs, il est intéressant d'examiner l'effet perturbateur pour le rang de monosémie

technique. Ci-dessus, nous avons évoqué l'effet perturbateur des 1507 spécificités générales, responsables de l'hétéroscédasticité, et l'homoscédasticité des 3210 spécificités restantes pour le rang de monosémie. Or, pour le rang de monosémie technique, l'effet perturbateur est encore plus important, parce que le sous-ensemble des 3210 spécificités se voit toujours confronté à l'hétéroscédasticité. L'exclusion d'un petit sous-ensemble supplémentaire s'impose donc ; au total 1594 mots doivent être exclus pour aboutir à l'homoscédasticité. Pour les 3123 spécificités restantes, le pourcentage de variation expliquée R^2 est d'ailleurs moins élevé (49,26%) pour le rang de monosémie technique que celui des 3210 spécificités pour le rang de monosémie (60,35%), ce qui indique une corrélation moins bonne pour le rang de monosémie technique. Cependant, pour le rang de monosémie technique, le pourcentage de variation expliquée R^2 pour le sous-ensemble homoscédastique des 3123 spécificités (49,26%) est supérieur au pourcentage pour l'ensemble des 4717 spécificités (42,74%). Lorsqu'on compare la droite de régression des 1507 mots exclus pour le rang de monosémie technique et celle des 1594 mots (Cf. annexe 13 : figure A13.1), on observe que les 1507 mots exclus sont plus polysémiques techniquement que les 1594 mots exclus. Autrement dit, si l'effet perturbateur des 1507 mots exclus est plus important pour le rang de monosémie technique, il n'est pas dû à une moindre polysémie technique.

7.1.5.2 Régression multiple : facteurs de fréquence et de recoupement

Afin de caractériser les deux sous-ensembles de 1507 et de 3210 spécificités, nous procédons à une analyse de régression multiple faisant intervenir comme variable dépendante (VD) le rang de monosémie (et le rang de monosémie technique) et comme variables indépendantes (VI) un certain nombre de facteurs importants en matière de fréquence et de recoupement des cc (Cf. chapitre 6 pour les 50 spécificités représentatives). Ces facteurs comprennent le nombre de longueurs de vecteurs-cc, le nombre moyen de vecteurs-cc par longueur, la longueur moyenne des vecteurs-cc, l'écart-type[178] des longueurs des vecteurs-cc, le recoupement moyen, le recoupement relatif moyen, le pourcentage de cc isolés, la fréquence moyenne des cc, l'écart-type des fréquences des cc, la technicité moyenne ces cc (ou la valeur de LLR moyenne des cc), l'écart-type des technicités des cc et, finalement, la fréquence moyenne pondérée (WLLR)[179] des cc.

[178] Si l'écart-type des longueurs des vecteurs-cc est élevé, les vecteurs-cc ont des longueurs très différentes. L'écart-type donne une idée de la distribution de la variation.

[179] Il s'agit de la fréquence de chaque cc (*cc-type*) pondérée par le facteur de pondération de sa technicité, utilisée dans la formule de la mesure de recoupement technique.

Après élimination des facteurs impliqués dans la multicolinéarité (Cf. 7.2.1 pour les détails techniques de la multicolinéarité), l'analyse de régression multiple pour les 1507 spécificités révèle que plusieurs facteurs sont significatifs (Cf. annexe 13). D'abord, plus le recoupement moyen et le recoupement relatif moyen sont élevés, plus le mot de base (ou la spécificité) est monosémique. En effet, plus les cc se recoupent en moyenne et même si on tient compte du nombre de cc par c, plus le mot de base est homogène sémantiquement. Ensuite, si le mot de base a plus de cc isolés (non partagés) et plus de cc techniques, il est plus monosémique, ce qui semble plutôt contradictoire, mais nous y reviendrons ci-après (Cf. 7.1.5.3). Finalement, plus il y a de vecteurs-cc par longueur, plus l'écart-type des longueurs des vecteurs-cc et l'écart-type des technicités[180] des cc sont importants, plus le mot de base est polysémique.

Cette analyse de régression multiple (avec tous les facteurs significatifs) fait émerger quelques corrélations bizarres et inattendues et mérite donc une mise au point à partir des corrélations individuelles. En plus, les analyses de régression multiple pour les 1507 et les 3210 spécificités aboutissent à des résultats assez similaires, tant pour le rang de monosémie (VD) que pour le rang de monosémie technique (VD) (Cf. annexe 13). Par conséquent, l'impact combiné de tous ces facteurs ne permet pas d'opposer clairement les deux sous-ensembles de spécificités et plaide pour une approche complémentaire à partir des corrélations individuelles des facteurs de fréquence et de recoupement des cc.

7.1.5.3 Corrélations et moyenne : facteurs de fréquence et de recoupement

Nous procédons donc pour les 1507 et les 3210 spécificités à la comparaison des corrélations individuelles avec, d'une part, chacun des facteurs de fréquence et de recoupement des cc et, d'autre part, le rang de monosémie et le rang de monosémie technique. Le document en annexe (Cf. annexe 13) visualise les résultats de toutes ces corrélations. Ici nous nous limitons aux corrélations les plus importantes pour l'analyse linguistique. A cet effet, nous ferons la distinction entre les facteurs linguistiques et les facteurs techniques, qui découlent de la formule de la mesure de recoupement.

- *Facteurs linguistiques*

Les facteurs linguistiques sont essentiellement axés sur le contenu interprétatif, parce qu'ils permettent de formuler des hypothèses interprétatives concernant la

---

[180] Les deux facteurs impliquant la technicité des cc (*technicité moyenne des cc* et *écart-type des technicités des cc*) n'ont pas de lien direct avec le rang de monosémie de base.

monosémie ou la polysémie plus ou moins grande à laquelle on peut s'attendre à partir de certains paramètres linguistiques intuitifs, tels que la fréquence moyenne des cc ou le pourcentage de cc isolés. Les facteurs linguistiques comprennent l'écart-type des longueurs des vecteurs-cc, l'écart-type des fréquences, le pourcentage de cc isolés, la technicité moyenne des cc et l'écart-type des technicités des cc. Les observations ci-dessous montreront que leurs corrélations avec le rang de monosémie et avec le rang de monosémie technique peuvent être peu intuitives, voire contre-intuitives. Les facteurs linguistiques et leurs corrélations sont donc interprétables jusqu'à un certain point et demandent à être complétés par d'autres facteurs, en particulier par des facteurs plus techniques.

– Ecart-type des longueurs des vecteurs-cc

Un premier facteur particulièrement intéressant pour l'opposition des deux sous-ensembles de 1507 et de 3210 spécificités, est l'écart-type des longueurs des vecteurs-cc (stdev_long). En effet, on pourrait avancer l'hypothèse que la percolation de la polysémie générale des 1507 spécificités se manifeste par une plus grande variation dans les longueurs des vecteurs-cc (stdev_long plus élevé). Ainsi, le fait d'avoir plus de longueurs différentes, c'est-à-dire le fait d'avoir plus de variation dans le nombre de cc par c, pourrait refléter la présence de plus de contextes (dont des contextes généraux) et entraîner des rangs de monosémie près de 4700. Or, les 1507 mots généraux plutôt polysémiques se caractérisent par des corrélations positives (de stdev_long), qui sont toutefois moins bonnes que les corrélations des 3210 spécificités. S'il est vrai que les 1507 mots généraux ont des rangs plutôt polysémiques, les corrélations moins bonnes (avec stdev_long) indiquent que ces 1507 mots n'ont, visiblement, pas plus de longueurs différentes des vecteurs-cc.

Toutefois, la moyenne de l'écart-type des longueurs des vecteurs-cc pour les 1507 et pour les 3210 spécificités, respectivement de 128,58 et de 71,53, indique que le groupe des 1507 spécificités a le plus de longueurs différentes. Ce sous-ensemble présente effectivement beaucoup plus de variation quant à la longueur des vecteurs-cc, mais celle-ci n'est pas proportionnelle aux rangs[181] de monosémie (technique) près de 4700, qui correspondent à des valeurs numériques très élevées. Dans le sous-ensemble des 3210 spécificités, la variation limitée des longueurs des vecteurs-cc est plus proportionnelle aux rangs de monosémie plus bas (plus monosémiques), d'où les meilleures corrélations pour ce sous-ensemble. Par conséquent, cette observation

[181] La qualité des corrélations avec les degrés de monosémie et de monosémie technique (valeurs très faibles, entre 0 et 1) est comparable à celle des corrélations avec les rangs de monosémie et de monosémie technique, bien que celles-ci soient négatives (Cf. annexe 13).

explique pourquoi les corrélations positives sont moins bonnes pour les 1507 spécificités que pour les 3210 spécificités. Elle confirme également l'hypothèse que la polysémie des 1507 spécificités se manifeste par la présence de plus de longueurs différentes des vecteurs-cc, même s'il ne s'agit pas seulement de polysémie générale, mais aussi de polysémie technique.

Comme les corrélations s'avèrent très sensibles aux valeurs ordinales à la fois élevées et hétérogènes (entre 1 et 4717) des rangs de monosémie et des rangs de monosémie technique, et que la moyenne permet de nuancer les résultats, nous avons décidé de déterminer la moyenne par sous-ensemble pour tous les facteurs de fréquence et de recoupement (Cf. annexe 13).

– Pourcentage de cc isolés (et cc uniques) + Ecart-type des fréquences moyennes

Le facteur linguistique du pourcentage de cc isolés exprime le pourcentage de cc non partagés ou le pourcentage de cc figurant une fois dans la liste des cc. En théorie, un pourcentage plus élevé de cc isolés, par rapport au nombre total de cc, représente moins de recoupement et correspond à des rangs de monosémie plus élevés (plus près de 4700). On s'attend dès lors à des corrélations positives pour le sous-ensemble des 1507 spécificités plutôt polysémiques. Néanmoins, les corrélations observées pour les 1507 spécificités sont négatives (-0,68 pour le rang de monosémie et -0,64 pour le rang de monosémie technique) et même plus importantes que celles des 3210 spécificités (-0,17 et -0,19) (Cf. annexe 13). En plus, la moyenne du pourcentage de cc isolés est plus faible (81%) pour les 1507 que pour les 3210 spécificités (89%), tandis que, en théorie, on s'attendrait à une moyenne plus élevée pour les 1507 spécificités plutôt polysémiques. De même, la moyenne du pourcentage de cc uniques[182] ou différents est plus faible (76%) pour les 1507 que pour les 3210 spécificités (87%) (Cf. annexe 13). Un pourcentage de cc uniques plus faible signifie moins de cc différents, c'est-à-dire plus de cc qui se recoupent et, par voie de conséquence, une homogénéité sémantique plus importante pour le mot de base. Or, les 1507 spécificités sont plutôt polysémiques, bien qu'elles se caractérisent par les pourcentages de cc uniques les plus faibles[183]. Par conséquent, les moyennes du pourcentage de cc isolés et de cc uniques confirment les corrélations négatives et contredisent l'interprétation intuitive.

---

[182] Le pourcentage de cc uniques correspond au nombre de cc uniques (*cc-types*) par rapport au nombre total de cc (*cc-tokens*).

[183] Les spécificités les plus polysémiques du sous-ensemble des 1507 spécificités, *machine* et *outil*, ont respectivement 38,9% et 46,2% de cc uniques par rapport au nombre total de cc.

Pour le facteur linguistique de l'écart-type des fréquences moyennes (stdev_fq), les 1507 spécificités se caractérisent par de meilleures corrélations positives et par un écart moyen plus élevé que les 3210 spécificités : plus les fréquences des cc sont différentes et élevées, plus les spécificités sont polysémiques. Cependant, en théorie, les fréquences élevées des cc indiquent un recoupement plus important.

Comment interpréter correctement ces facteurs linguistiques, les corrélations et les moyennes ? D'une part, il est à noter que les corrélations n'expriment pas nécessairement une relation de cause à effet. D'autre part, le calcul du recoupement, tel qu'il est implémenté dans la mesure de monosémie et de monosémie technique, ne s'appuie pas sur le pourcentage de cc uniques, ni sur le pourcentage de cc isolés, mais sur les cc partagés et plus particulièrement, sur la répartition des cc partagés (Cf. chapitres 5 et 6). En plus, il tient compte du nombre total de c et de cc. Les corrélations négatives importantes pour le pourcentage de cc isolés (ou de cc uniques), de même que les corrélations positives importantes pour l'écart-type des fréquences moyennes devront donc être compensées par d'autres facteurs. S'il y a plus de cc plus fréquents (partagés) ou moins de cc isolés, on peut se demander quel est le nombre total de cc par mot de base et comment les cc partagés sont répartis parmi les c.

En effet, il est important de connaître également la répartition des cc ou le recoupement relatif moyen (Cf. ci-dessous), ainsi que le nombre total de cc. La visualisation ci-dessous (Cf. figure 7.14) montre non seulement une corrélation négative[184] entre le nombre total de cc (c_0.9999) et le pourcentage de cc uniques (perc_ccuni), mais également une augmentation exponentielle du nombre total de cc pour les spécificités ayant un pourcentage plus faible de cc uniques[185] (à gauche). Qui plus est, pour le même pourcentage de cc uniques, par exemple 0.7, les mots les plus polysémiques (en rouge) ont le nombre total de cc le plus élevé. Par conséquent, le nombre total de cc a plus d'impact sur l'hétérogénéité sémantique que le pourcentage de cc isolés ou de cc uniques.

---

[184] Le coefficient de corrélation (Pearson) entre le nombre total de cc et le pourcentage de cc uniques correspond à -0,89. Il en va de même pour le pourcentage de cc isolés (-0,86).

[185] L'explication statistique indique que, si l'espace de recherche est plus limité (moins de cc au total), la probabilité de trouver des cc identiques est plus faible et donc le pourcentage de cc uniques sera plus élevé. En revanche, l'explication linguistique précise que, plus le mot est fréquent, plus il a de cc au total (corrélation positive) et plus grande sera la chance qu'on trouve le même cc (Cf. *Type-Token-Ratio*), mais aussi que le mot revête plusieurs sens (explication ambiguë).

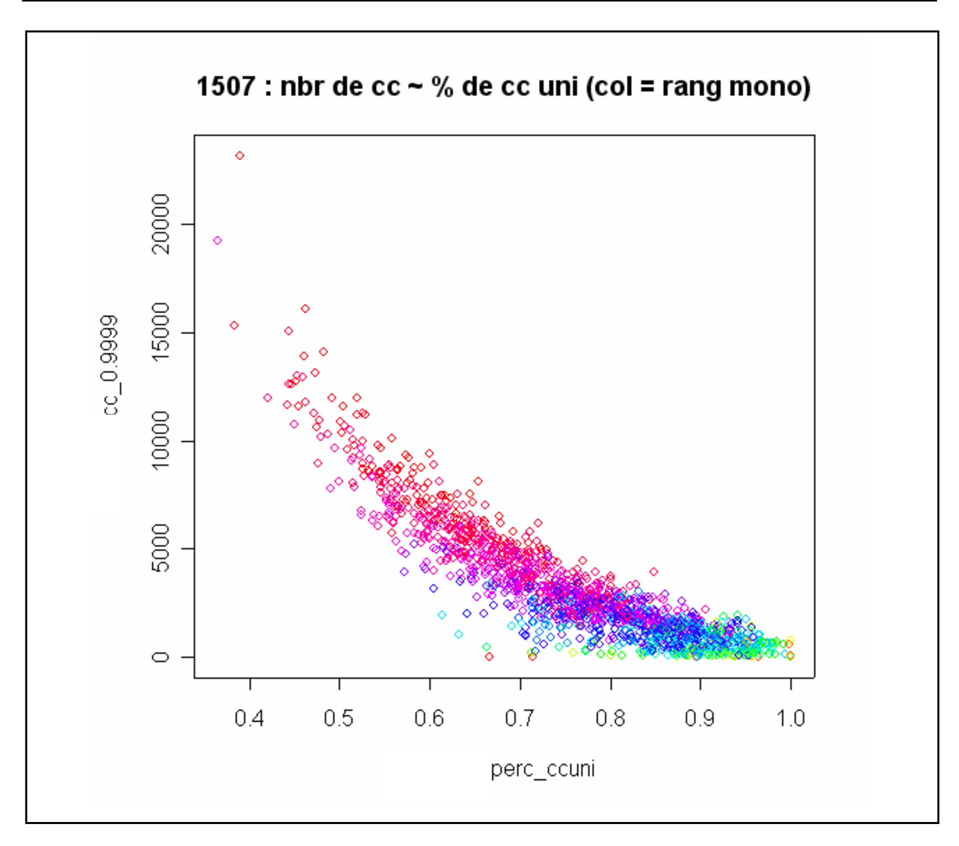


Figure 7.14 Sous-ensemble des 1507 spécificités : nombre total de cc ~ pourcentage de cc uniques (rang de monosémie en couleur)

En fait, on pourrait imaginer une formule plutôt simple et « naïve » d'homogénéité sémantique, à partir du pourcentage de cc uniques. Un nombre plus élevé de cc uniques (*cc-types*) par rapport au nombre total de cc (*cc-tokens*) signifierait moins de recoupement des cc-types et donc plus d'hétérogénéité sémantique. Par contre, un pourcentage plus limité de cc uniques indiquerait plus de recoupement des cc-types et donc plus d'homogénéité sémantique. Selon cette formule naïve, basée sur le pourcentage de cc uniques, les spécificités avec le moins de cc-types par rapport au nombre total de cc, telles que *machine* et *outil*, seraient les plus homogènes sémantiquement et, en revanche, les spécificités avec le pourcentage le plus élevé de cc-types seraient les plus hétérogènes sémantiquement. Comme ces spécificités, peu fréquentes et ayant (très) peu de cc au total, pourraient éventuellement relever de plusieurs domaines (Cf. nomadisation), cela permettrait d'expliquer leur hétérogénéité sémantique. Toutefois, la vérification de cette hypothèse requiert beaucoup plus de contextes, qui relèveraient de plusieurs domaines connexes, et qui contiendraient donc beaucoup plus de cc au total.

Les observations concernant la formule naïve plaident donc en faveur d'une formule d'homogénéité sémantique qui intègre plusieurs facteurs, en particulier le nombre total de cc. En effet, à gauche de la visualisation ci-dessus (Cf. figure 7.14), les spécificités ont beaucoup de cc au total et un pourcentage limité de cc uniques : elles sont hétérogènes sémantiquement en dépit de leur pourcentage limité de cc uniques. A droite, les spécificités ont très peu de cc au total, elles sont homogènes sémantiquement (indiquées en vert et en jaune), malgré leur pourcentage élevé de cc uniques.

Il est à noter que notre formule de monosémie ne permet pas de détecter la polysémie éventuelle des mots peu fréquents ayant (très) peu de cc au total. Le problème de la rareté des données au niveau du nombre total de cc s'accompagne en plus d'une ambiguïté interprétative : si les mots peu fréquents sont probablement plus homogènes sémantiquement, ils ont un pourcentage plus limité de cc partagés. La solution méthodologique la plus efficace réside donc dans la prise en compte d'un nombre suffisamment important de cc par mot de base.

– Technicité moyenne des cc + Ecart-type des technicités des cc

Avant de passer au facteur technique de recoupement, nous examinerons deux facteurs linguistiques portant sur la technicité des cc, c'est-à-dire sur le degré de spécificité des cc dans le corpus technique. Comme nous opposons un sous-ensemble de 1507 spécificités générales à un sous-ensemble de 3210 spécificités plutôt techniques, la technicité des cc constitue une piste de recherche intéressante. On pourrait argumenter que les cc des 1507 spécificités générales se caractérisent par une technicité[186] moyenne plus faible, parce qu'elles ont probablement plus de cc généraux. Dès lors, l'écart-type des technicités des cc serait plus important pour ce sous-ensemble. La différence de technicité qu'il y aurait entre les sous-ensembles devrait se manifester à travers les corrélations avec le rang de monosémie technique, parce que le degré de monosémie technique est calculé en fonction de la technicité des cc. Ces deux facteurs de technicité semblent donc importants pour la distinction des deux sous-ensembles. D'autant plus que l'analyse de régression multiple qui fait intervenir les facteurs de fréquence et de recoupement (Cf. 7.1.5.2), a démontré la pertinence de la technicité moyenne des cc et de l'écart-type des technicités des cc, tant pour le rang de monosémie technique que pour le rang de monosémie (de base).

---

[186] Plus la technicité moyenne des cc est élevée, plus les cc sont techniques (globalement) et plus ils pèsent lourd dans le recoupement technique (entraînant une monosémie technique).

Toutefois, les corrélations de ces deux facteurs avec le rang de monosémie technique ne permettent pas de confirmer les hypothèses formulées ci-dessus. En effet, les corrélations individuelles sont très faibles[187] : il n'y a donc pas de corrélation, même si ces facteurs contribuent légèrement au modèle de régression multiple. Les moyennes cependant sont plus intéressantes. Les cc des 1507 spécificités ont une technicité moyenne plus élevée (560) que ceux des 3210 spécificités (524) : ils sont donc plus techniques (plus spécifiques du corpus technique), contrairement à l'hypothèse formulée pour les 1507 mots généraux. Cette technicité moyenne plus élevée des mots généraux s'explique principalement par le fait qu'un nombre important de mots généraux sont des mots plutôt spécifiques (à gauche), tels que *machine* et *outil*, avec quelques cc très techniques[188] qui augmentent considérablement la moyenne. En ce qui concerne l'écart-type des technicités, le sous-ensemble des 1507 spécificités générales a une moyenne plus élevée (2001) que l'autre sous-ensemble (1740). Un écart-type plus élevé est révélateur de plus de technicités (valeurs de LLR des cc) différentes, tant élevées que faibles. Si les 1507 spécificités générales ont globalement beaucoup de cc très techniques, certaines spécificités parmi les 1507 spécificités générales ont tout de même beaucoup de cc non techniques également, cela étant vrai surtout des spécificités les moins spécifiques (à droite), telles que *service, objet, commercial*.

Etant donné que les cc techniques entraînent des rangs de monosémie technique plus bas et que les cc moins techniques sont responsables des rangs de monosémie technique plus élevés, les moyennes de ces deux facteurs permettent d'expliquer les observations formulées ci-dessus pour la monosémie technique des 1507 spécificités générales. Les mots les plus spécifiques, ayant le plus de cc techniques, deviennent un peu plus monosémiques techniquement. Les mots les moins spécifiques, ayant le moins de cc techniques, deviennent un peu plus polysémiques techniquement.

Comme la technicité des cc entraîne une évolution différente en fonction du rang de spécificité et que nous voulons caractériser qualitativement l'effet de la technicité des cc, nous procédons finalement à l'identification des spécificités les plus

---

[187] La technicité moyenne se caractérise par un coefficient de corrélation de -0,05 pour les 1507 spécificités et de 0,05 pour les 3210 spécificités. L'écart-type des technicités moyennes a un coefficient de corrélation de 0,02 pour les 1507 spécificités et de 0,19 pour les 3210 spécificités.

[188] Ces spécificités très spécifiques, très fréquentes et polysémiques, telles que *machine* et *outil*, entrent souvent dans la composition d'unités polylexicales. Ces unités polylexicales étant considérées comme des unités terminologiques, elles s'accompagnent souvent de cooccurrents (cc du mot de base) plutôt techniques.

sensibles à la technicité des cc, c'est-à-dire à l'effet de la mesure de monosémie technique. Les 219 mots ayant la plus grande différence[189] entre le degré de monosémie et le degré de monosémie technique se situent majoritairement (90%) dans le sous-ensemble des 3210 spécificités techniques. Ces 219 mots sensibles (Cf. annexe 13 pour les détails : 13.6) sont des mots peu spécifiques et plutôt monosémiques (ils se trouvent à droite en bas). Ils deviennent beaucoup moins monosémiques techniquement (à droite, plus en haut), principalement en raison de la faible technicité moyenne des cc (moyenne de liste de 146) et en raison du nombre limité de cc (moyenne de liste de 169). Ces mots se caractérisent donc par une monosémie générale (compte tenu de tous les cc)[190], en dépit de leur pourcentage important de cc isolés (moyenne de 86%). Les mots avec peu de cc au total se situent dans la zone peu fiable (Cf. ci-dessus), qui s'avère en plus être la zone la plus sensible à l'effet de technicité. Parmi ces 219 mots, on retrouve entre autres les spécificités (peu spécifiques et peu centrales par rapport au domaine technique) *télécom, fondamentalement, codage, sous-tendre, socio-économique, excavateur*.

Par contre, les 119 mots qui ont la différence la plus faible entre le degré de monosémie et le degré de monosémie technique, se caractérisent par des cc très techniques (moyenne de liste de 723) et par un nombre très important de cc au total (moyenne de liste de 3298 cc). Ces mots deviennent un peu moins polysémiques techniquement, principalement en raison du nombre important de cc techniques, qui ont en plus une technicité moyenne élevée. La présence massive de cc techniques a un effet significatif sur le degré de monosémie technique. Ces 119 spécificités comprennent notamment *reconditionnement, nitrurer, broche, mm, t/mn, numérique*.

- *Facteurs techniques*

Comme nous l'avons évoqué ci-dessus, l'interprétation des facteurs linguistiques devrait aussi prendre en considération le nombre de cc et le mode de répartition des cc partagés. Une interprétation plus adéquate des corrélations et des moyennes s'appuie donc sur des facteurs plus techniques, qui découlent de la formule de recoupement (technique), tels que la fréquence moyenne pondérée et le recoupement relatif moyen.

---

[189] Le degré de monosémie technique divisé par le degré de monosémie (ici inférieur à 0,50) donne une bonne idée du rapport, parce que le degré de monosémie technique est toujours inférieur au degré de monosémie.

[190] Les cc généraux sont responsables du recoupement, par exemple pour *télécom* les cc généraux *télécoms* et *collectivités*, pour *codage* les cc généraux *codage* et *clichés* et , .

– Fréquence moyenne pondérée

Parmi les facteurs plutôt techniques, la fréquence moyenne pondérée (fq_moy_wllr) constitue un facteur très intéressant du point de vue de l'opposition des deux sous-ensembles, parce qu'il combine la technicité des cc et la fréquence des cc (donc leur recoupement). Rappelons qu'une fréquence moyenne pondérée élevée signifie plus de cc partagés (plus de recoupement), qui sont en outre plus techniques. Le facteur donne donc également une idée de la richesse en cc techniques.

Les 1507 spécificités ont une moyenne plus élevée pour la fréquence moyenne et pour la fréquence moyenne pondérée, ce qui indique qu'elles ont plus de cc partagés et plus de cc techniques partagés que les 3210 spécificités. Cette observation pourrait signaler des rangs de monosémie technique plus bas, si le nombre total de cc des spécificités des deux sous-ensembles était équivalent. Or, les 1507 spécificités générales ont beaucoup plus de cc au total que les 3210 spécificités techniques. En plus, les corrélations sont positives et même plus importantes pour les 1507 spécificités (0,63 et 0,52) que pour les 3210 spécificités (0,42 et 0,22) : une fréquence moyenne pondérée plus élevée s'accompagne de rangs de monosémie plus élevés, plus particulièrement pour les 1507 spécificités générales. Ces corrélations positives sont tout à fait justifiées, parce que les 1507 spécificités sont plus polysémiques, bien que contraires à la logique de la fréquence élevée des cc (recoupement). Il faut nuancer la fréquence élevée des cc de ces spécificités, en tenant compte du recoupement relatif moyen de leurs cc.

Comme la fréquence moyenne pondérée équivaut au numérateur de la formule pour le recoupement technique, il est clair que l'interprétation correcte nécessite la prise en compte d'autres facteurs, tels que le nombre total de c et de cc et la façon dont les cc sont partagés, à savoir le recoupement (relatif) moyen.

– Recoupement relatif moyen

Comme nous l'avons expliqué ci-dessus (Cf. chapitre 6), le recoupement relatif moyen des cc tient compte du nombre de cc dans les couples de vecteurs-cc qui sont comparés pour déterminer le recoupement moyen. Nous observons des corrélations négatives entre le recoupement relatif moyen et les rangs de monosémie et de monosémie technique, un peu moins fortes pour les 1507 spécificités (-0,35 et -0,32) que pour les 3210 spécificités (-0,45 et -0,42). Plus le recoupement relatif moyen est élevé, plus le mot de base est monosémique, ce qui est parfaitement logique et intuitif. Par ailleurs, les moyennes indiquent que le recoupement relatif moyen des 1507 spécificités (0,047) est inférieure à celui des 3210 spécificités (0,107) et confirment dès lors la polysémie (technique) plus importante des 1507 spécificités générales.

- *Interprétation des facteurs linguistiques et techniques*

Dans les analyses des corrélations et des moyennes des facteurs de fréquence et de recoupement des 1507 et des 3210 spécificités nous avons fait appel à des facteurs linguistiques et des facteurs d'ordre technique. Bien que les facteurs linguistiques, axés sur le contenu interprétatif, permettent de formuler des hypothèses linguistiques, ils demandent à être compensés par des facteurs plus techniques. Grâce à la prise en compte de tous ces facteurs, nous avons pu interpréter les données et distinguer les deux sous-ensembles de spécificités. Les deux facteurs distinctifs les plus importants sont la fréquence moyenne pondérée (qui intègre la fréquence (recoupement) et la technicité des cc) et le recoupement relatif moyen (qui intègre le recoupement des cc et le nombre total de cc).

Si on reprend les deux axes de la visualisation de base, l'axe X représente le rang de spécificité et l'axe Y représente le rang de monosémie ou le rang de monosémie technique. On constate que ces deux facteurs évoluent chacun en fonction d'un axe (Cf. figure 7.15). Compte tenu de la fréquence moyenne pondérée et du recoupement relatif moyen, qui tous les deux peuvent être, selon le cas, importants ou faibles, nous pouvons effectuer une comparaison croisée et distinguer quatre cas de figure dans la visualisation de base :

1)	+	+	en bas à gauche :	presque vide
2)	+	-	en haut à gauche :	mots spécifiques et polysémiques
3)	-	+	en bas à droite :	mots peu spécifiques et monosémiques
4)	-	-	en haut à droite :	mots peu spécifiques et polysémiques

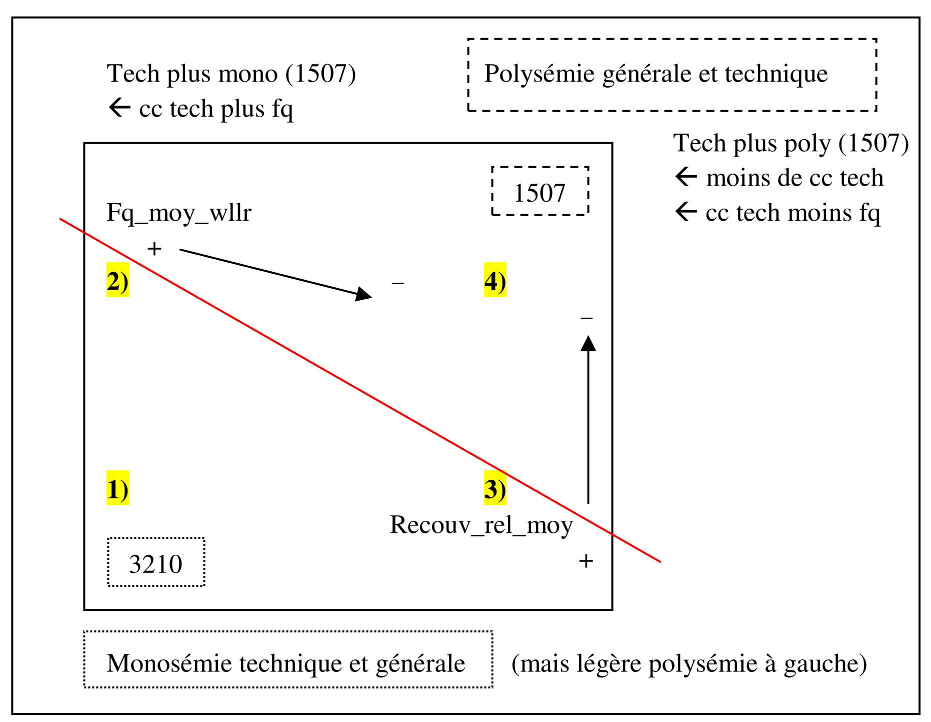


Figure 7.15 Fréquence moyenne pondérée et recoupement relatif moyen

Comme le montre la figure 7.15, la fréquence moyenne pondérée diminue en fonction de la moindre spécificité du mot de base (rangs de spécificité plus élevés ou plus près de 4700), c'est-à-dire de gauche à droite et aussi légèrement de haut en bas (rangs de monosémie plus bas). On se rappellera ici les corrélations positives entre la fréquence moyenne pondérée et les rangs de monosémie et de monosémie technique. Par contre, la corrélation entre la fréquence moyenne pondérée et le rang de spécificité est négative (Cf. annexe 13). Le recoupement relatif moyen en revanche diminue au fur et à mesure que les rangs de monosémie ou de monosémie technique sont plus élevés. Le recoupement relatif moyen diminue donc du bas vers le haut, ce qui est signalé par les corrélations négatives entre le recoupement relatif moyen et les rangs de monosémie et de monosémie technique.

En conclusion, il est clair que le recoupement relatif moyen est un facteur plus puissant que la fréquence moyenne pondérée. Le recoupement relatif moyen plus bas en haut de la visualisation est tellement fort qu'il compense à gauche, pour les mots les plus spécifiques, la fréquence de leurs cc techniques qui se recoupent. S'il est vrai que ces spécificités générales (bien que spécifiques) ont beaucoup de cc fréquents, le recoupement relatif de ces cc n'est pas si important, car elles ont énormément de cc au total, ce qui rend la fréquence relative des cc moins importante.

7.1.6 Conclusion pour les 3210 spécificités techniques

Les expérimentations de répartition et d'exclusion, visant principalement à résoudre le problème de l'hétéroscédasticité des 4717 spécificités, ont contribué également et surtout à une meilleure compréhension des caractéristiques des spécificités. Ainsi, nous avons identifié deux sous-ensembles de spécificités : un sous-ensemble de 1507 spécificités plutôt générales et un sous-ensemble de 3210 spécificités techniques. Les 1507 spécificités sont des mots généraux et se caractérisent par une polysémie générale et technique qui entraîne un effet perturbateur par rapport à la tendance générale. Celle-ci correspond à une corrélation négative entre, d'une part, le rang de spécificité et, de l'autre, le rang de monosémie et le rang de monosémie technique.

L'exclusion des 1507 mots généraux nous a permis d'isoler 3210 spécificités techniques qui se caractérisent par l'homoscédasticité et par un pourcentage de variation expliquée R^2 élevé (60,35%), c'est-à-dire par une corrélation linéaire négative entre le rang de spécificité et le rang de monosémie[191]. Ces 3210 spécificités peuvent être plus ou moins spécifiques du corpus technique et plus ou moins fréquentes dans le corpus technique, mais elles sont en tout cas très peu fréquentes ou même absentes du corpus de référence de langue générale. Leur variation quant au rang de spécificité permet de rendre compte et de prédire la variation quant au rang de monosémie. Il s'ensuit que, parmi les 3210 spécificités, les mots les plus spécifiques de notre corpus technique (*usinage, broche, arête, découpe*) sont plutôt polysémiques et que les mots les moins spécifiques (*infalsifiable, adhésif, présentoir, transmissible*) sont plutôt monosémiques[192]. Cette conclusion générale s'oppose clairement à la thèse monosémiste, comme nous l'avons signalé à plusieurs reprises.

Les 3210 spécificités techniques se caractérisent en gros par un recoupement relatif moyen très important, ce qui explique leur homogénéité sémantique considérable. Elles présentent globalement peu de variation dans les fréquences, dans les technicités des cc et dans les longueurs des vecteurs-cc, avec des cc moins fréquents

---

[191] Pour le rang de monosémie technique, nous recensons 3123 spécificités pour obtenir l'homoscédasticité.

[192] Signalons à ce sujet l'importance du domaine. Cette conclusion n'est valable que pour notre corpus technique, qui relève du domaine des machines-outils pour l'usinage des métaux. Si on conduit l'analyse des spécificités sur un autre corpus spécialisé, on trouvera d'autres mots spécifiques, qui seront représentatifs du domaine en question (et non pas *machine, outil, usinage, broche, pièce*, etc.). Probablement, les mots les plus spécifiques du nouveau corpus seront également plutôt polysémiques.

et moins techniques et avec moins de cc par c. En plus, le nombre total de cc par mot de base, ainsi que le pourcentage de cc isolés sont plutôt faibles. Donc, leur pourcentage de cc partagés est plus important. Or, il faut le compenser immédiatement par le recoupement relatif moyen, peu important, et par le nombre total de cc, globalement peu important. Même si les spécificités les plus spécifiques parmi les 3210 sont tout de même assez polysémiques, cela s'explique principalement par le fait que leur fréquence moyenne pondérée est assez importante (cc techniques fréquents) et que leur nombre de cc est très élevé, d'ailleurs plus élevé que le nombre moyen de cc des 3210 spécificités.

Finalement, il est important de souligner que l'exclusion des 1507 spécificités très générales, apparemment trop générales pour le bon fonctionnement du modèle, n'a pas été effectuée parce que nous voulions obtenir à tout prix une corrélation linéaire. Si nous avons isolé les spécificités qui confirment la tendance générale et partant, la puissance explicative du modèle, tout en excluant les autres, c'est parce que nous voulions cerner de plus près les caractéristiques linguistiques des deux groupes et interpréter correctement le modèle statistique de régression simple. Il est clair que les mots très fréquents dans le corpus général ne se prêtent pas à une prédiction de leur rang de monosémie à partir de leur rang de spécificité, car ils restent plutôt polysémiques, quel que soit leur rang de spécificité.

En guise de conclusion, le modèle statistique de régression linéaire simple n'est guère satisfaisant pour des mots généraux, qui sont fréquents dans un corpus de langue générale, mais qui s'avèrent quand même spécifiques du corpus technique, en raison de leur fréquence technique très élevée. On peut dès lors se poser la question de savoir quel est l'impact sur les rangs de monosémie et de monosémie technique, si le modèle de régression comprend plusieurs variables indépendantes, telles que la fréquence générale et la fréquence technique ? La régression multiple fera l'objet de la deuxième partie de ce chapitre (Cf. 7.2).

7.2 ANALYSE DE RÉGRESSION MULTIPLE

Etant donné que le rang de monosémie et le rang de monosémie technique ne sont pas uniquement influencés par le rang de spécificité, mais également par d'autres variables, nous procédons à une analyse statistique de régression multiple pour les 4717 spécificités (Cf. Bertels, Speelman & Geeraerts 2006). L'analyse de régression multiple fera intervenir toutes les variables indépendantes susceptibles d'influer sur la variable dépendante, c'est-à-dire le rang de monosémie ou le rang de monosémie technique des 4717 spécificités.

Les variables indépendantes sont principalement des variables quantitatives, tant numériques (p.ex. fréquence, longueur) qu'ordinales (p.ex. rang). Les variables indépendantes comprennent non seulement le rang de spécificité des 4717 spécificités, mais également le degré de spécificité (log du LLR)[193], le rang de fréquence dans le corpus technique et dans le corpus général, la fréquence absolue dans le corpus technique et dans le corpus général, la longueur (nombre de caractères), ainsi que la classe lexicale et le nombre de classes lexicales. Afin de déterminer la classe lexicale (unique ou prédominante) des 4717 spécificités, nous recourons aux fichiers lemmatisés de l'analyseur Cordial, qui comprennent, outre la forme graphique et la forme lemmatisée, le code Cordial qui indique la classe lexicale[194]. Ainsi, deux variables indépendantes supplémentaires se rajoutent par spécificité : la variable quantitative numérique du nombre de classes lexicales différentes auxquelles elle appartient (de 1 à 4) et la variable qualitative catégorielle de la classe lexicale (*adj, adv, nom, verbe, func* ou *nprop*[195]).

Dans cette deuxième partie, nous formulerons d'abord une mise en garde par rapport à la multicolinéarité[196] (7.2.1). Ensuite, nous présenterons les résultats de diverses analyses de régression, en fonction de plusieurs combinaisons des variables indépendantes (7.2.2). Nous terminerons le chapitre par une conclusion (7.2.3).

7.2.1 Le problème de la multicolinéarité

Le but de l'analyse de régression multiple est d'évaluer l'impact combiné et simultané de plusieurs variables indépendantes sur la variable dépendante, en l'occurrence le rang de monosémie et, dans un deuxième temps, le rang de monosémie technique. Ces variables indépendantes ou explicatives servent à prédire la variation de la variable dépendante. Malheureusement, les variables indépendantes du modèle de régression multiple ne sont pas toujours indépendantes

---

[193] Rappelons que le log du degré de spécificité (log_LLR) permet de rééchelonner les degrés de spécificité ou valeurs de LLR (de 50521 à 3,85) entre 4,70 et 0,58.

[194] Les détails de ces opérations sont expliqués dans le document en annexe (Cf. annexe 14).

[195] Les valeurs *func* et *nprop*, respectivement « mots grammaticaux » et « noms propres », correspondent à des spécificités qui ont plusieurs codes de plusieurs classes lexicales différentes, mais dont les codes *func* et *nprop* sont les plus fréquents. Rappelons que certains noms propres ont été maintenus (*Cao, Cnc, Cfao*, …), parce qu'il s'agit de sigles importants.

[196] Par analogie avec *colinéaire*, nous adoptons l'orthographe *multicolinéarité*. Notons que le glossaire de termes statistiques ISI (*International Statistical Institute*) écrit *multicollinéarité* (Cf. http://europa.eu.int/en/comm/eurostat/research/isi/concepts/concept01907.htm).

les unes des autres. Parfois, deux ou plusieurs variables indépendantes sont corrélées entre elles, autrement dit, elles expliquent en grande partie la même variation de la variable dépendante. C'est le problème de la multicolinéarité : plusieurs variables sont « colinéaires ». Il est important de vérifier la multicolinéarité des variables avant de passer à l'analyse de régression multiple, car elle entraîne deux conséquences méthodologiques.

7.2.1.1 Conséquences de la multicolinéarité

Tout d'abord, la multicolinéarité mène à une augmentation des écarts-types des estimations de coefficient dans le modèle de régression multiple. Par conséquent, on trouvera moins vite des rapports significatifs entre les variables indépendantes et la variable dépendante. Lorsqu'on procède à des tests t pour déterminer la significativité des coefficients particuliers, on risque de trouver qu'aucune des variables indépendantes n'est significative, tandis que le test F du modèle de régression multiple révèle une significativité importante. En plus, la multicolinéarité rend le modèle de régression multiple peu fiable, parce qu'elle accroît l'erreur sur les valeurs estimées de la variable dépendante. Compte tenu de ces deux problèmes, il importe de détecter la multicolinéarité et de la résoudre, avant de passer à l'analyse de régression multiple.

7.2.1.2 La solution : le calcul des VIF

Pour détecter des problèmes de multicolinéarité lorsqu'on fait intervenir deux ou plusieurs variables indépendantes, on fait appel au facteur d'inflation de la variance (VIF ou *Variance Inflation Factor*). On calcule le VIF d'une variable indépendante en considérant cette variable comme variable dépendante d'une analyse de régression multiple particulière avec toutes les autres variables indépendantes comme variables indépendantes. Si cette variable est caractérisée par des rapports linéaires avec les autres variables, son coefficient de détermination (R^2) ou pourcentage de variation expliquée sera élevé.

Le calcul des VIF est implémenté dans R et se fait simultanément pour toutes les variables indépendantes d'un modèle de régression multiple. Il est à noter que la variable catégorielle (la classe lexicale) sera exclue de cette vérification des VIF[197]. Un VIF supérieur à 10 (Welkenhuysen-Gybels & Loosveldt 2002) signale un problème de multicolinéarité et, le cas échéant, toutes les variables impliquées dans le rapport colinéaire auront un VIF très (ou trop) élevé. La solution du problème de

---

[197] Le calcul des facteurs d'inflation de la variance ou des VIF prend en considération uniquement des variables numériques, donc pas des variables catégorielles : $VIF = 1/(1-R^2)$.

multicolinéarité consiste à exclure du modèle de régression multiple une des variables indépendantes avec un VIF trop élevé, en l'occurrence celle avec le VIF le plus élevé. Cette procédure est réitérée jusqu'à ce que toutes les variables indépendantes impliquées aient un VIF inférieur à 10 et puissent être intégrées dans le modèle de régression multiple.

La matrice des corrélations (Cf. annexe 14) montre un coefficient de corrélation Pearson trop élevé (supérieur à 0,90) entre le rang de spécificité et le log du LLR. Ces deux variables sont clairement intercorrélées, étant donné que les rangs de spécificité sont attribués à partir du classement des degrés de spécificité (valeurs de LLR). Le calcul des VIF ci-dessous signale effectivement un problème de multicolinéarité pour trois variables : le log du LLR (VIF 36,26), le rang de spécificité (VIF 26,32) et le rang de fréquence technique (VIF 14,72) (Cf. tableau 7.15). Deux options sont possibles : (a) la suppression du log_LLR, qui a le VIF le plus élevé, (b) la suppression du rang de spécificité. Dans le dernier cas, on peut maintenir le log_LLR en raison de son coefficient de corrélation un peu plus élevé avec la variable dépendante (Cf. matrice des corrélations : annexe 14). Notons que la suppression du log du LLR ne permet pas de résoudre tout le problème de multicolinéarité, parce que le VIF du rang de fréquence technique reste toujours trop élevé (Cf. tableau 7.15), ce qui vaut également pour la suppression du rang de spécificité.

```
> resM <- ols(rang_v_mono_0.9999 ~ rang_v_spec + log_LLR + rang_v_freq1 +
rang_v_freq2 + freqabs1 + freqabs2 + nbr_claslex + long, data = m)
> vif(resM)
 rang_v_spec      log_LLR rang_v_freq1 rang_v_freq2     freqabs1     freqabs2
   26.326119    36.269684    14.727362     6.188937     3.624532     1.994256
 nbr_claslex         long
    1.070609     1.095690

> resM <- ols(rang_v_mono_0.9999 ~ rang_v_spec + rang_v_freq1 + rang_v_freq2 +
freqabs1 + freqabs2 + nbr_claslex + long, data = m)
> vif(resM)
 rang_v_spec rang_v_freq1 rang_v_freq2     freqabs1     freqabs2  nbr_claslex
    5.805897    12.764813     5.655836     1.711915     1.548450     1.070074
        long
    1.094107
```

Tableau 7.15 Calcul des VIF pour toutes les variables indépendantes

En raison de la corrélation très importante entre le rang de fréquence technique et la variable dépendante (rang de monosémie et rang de monosémie technique), nous avons préféré garder le rang de fréquence technique comme variable indépendante. Sa corrélation importante avec le rang de fréquence générale (Cf. annexe 14), permet de supprimer celui-ci. Par conséquent, la multicolinéarité est résolue et le rang de fréquence technique est maintenu. Afin de maintenir tout de même la différence (ou l'écart) entre le rang de fréquence générale et le rang de fréquence technique, nous envisageons d'intégrer dans nos recherches futures une variable indépendante supplémentaire, à savoir l'écart des rangs de fréquence (Cf. 7.1.4.2).

Si le rang de fréquence générale est supprimé, cette nouvelle variable permettra de reprendre partiellement l'information perdue, sans que se pose le problème de multicolinéarité (Cf. tableau 7.16).

```
> resM <- ols(rang_v_mono_0.9999 ~ rang_v_spec + log_LLR + rang_v_freq1 +
ecart_r_v_freq + freqabs1 + freqabs2 + nbr_claslex + long, data = m)
> vif(resM)
   rang_v_spec        log_LLR   rang_v_freq1 ecart_r_v_freq       freqabs1
     26.326119      36.269684       4.337908       2.673793       3.624532
      freqabs2    nbr_claslex           long
      1.994256       1.070609       1.095690

> resM <- ols(rang_v_mono_0.9999 ~ rang_v_spec + rang_v_freq1 + ecart_r_v_freq +
freqabs1 + freqabs2 + nbr_claslex + long, data = m)
> vif(resM)
   rang_v_spec   rang_v_freq1 ecart_r_v_freq       freqabs1       freqabs2
      5.805897       3.805852       2.443478       1.711915       1.548450
   nbr_claslex           long
      1.070074       1.094107
```

Tableau 7.16 Calcul des VIF avec l'écart des rangs de fréquence

Plusieurs analyses de régression multiple sont à envisager, tant pour le rang de monosémie que pour le rang de monosémie technique, et cela en fonction de plusieurs possibilités d'intégration des variables indépendantes : (1) la suppression de la variable avec un VIF trop élevé, soit (a) le log du LLR, soit (b) le rang de spécificité, (2) le choix d'intégrer ou non la variable combinée (log du LLR et écart des rangs de fréquence) et (3) le choix d'intégrer ou non la variable catégorielle de la classe lexicale. Il est à noter que le calcul des VIF des variables indépendantes vaut tant pour le rang de monosémie que pour le rang de monosémie technique comme variable dépendante, étant donné que les variables indépendantes du calcul des VIF sont les mêmes pour les deux.

7.2.2 Résultats de l'analyse de régression multiple

Dans cette section, nous procéderons à plusieurs analyses de régression multiple, principalement pour le rang de monosémie (7.2.2.1) et pour le rang de monosémie technique (7.2.2.2). Les choix explicités ci-dessus mèneront finalement à des analyses de régression multiple qui font intervenir d'autres configurations des variables indépendantes (7.2.2.3).

7.2.2.1 Le rang de monosémie

- *Maintien du rang de spécificité*

L'analyse de régression multiple principale est celle qui prend comme variable dépendante le rang de monosémie (de base) des 4717 spécificités et qui supprime, après le calcul des VIF, le log du LLR. La variable indépendante de l'analyse de régression simple, à savoir le rang de spécificité (rang_v_spec), est donc maintenue

dans le modèle d'analyse multiple. Dans cette analyse, le rang de fréquence générale est remplacé par l'écart des rangs de fréquence.

Après vérification des VIF et après avois effectué les choix méthodologiques commentés ci-dessus, nous avons procédé à une analyse de régression multiple « pas à pas » (*stepwise multiple regression*). Dans R, le modèle de régression multiple commence par toutes les variables indépendantes intégrées. Par défaut, il supprime automatiquement les variables indépendantes non significatives, par ordre décroissant de valeur p.

Les variables indépendantes significatives (Cf. tableau 7.17) expliquent 80,65% de la variation du rang de monosémie, à savoir le rang de fréquence technique, le rang de spécificité, la longueur et le nombre de classes lexicales. Même si la fréquence absolue dans le corpus technique (freqabs1) n'est pas significative, la régression multiple pas à pas maintient cette variable en raison de son apport au modèle (en termes de R² ou de statistique F).

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4609.38681   43.74092 105.379  < 2e-16 ***
rang_v_spec    -0.07575    0.01032  -7.343 2.46e-13 ***
rang_v_freq1   -0.85618    0.01121 -76.347  < 2e-16 ***
long          -20.18410    2.74861  -7.343 2.44e-13 ***
nbr_claslex    66.03865   23.52732   2.807  0.00502 **
freqabs1        0.03242    0.02139   1.516  0.12961
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 599.8 on 4711 degrees of freedom
Multiple R-Squared: 0.8067,     Adjusted R-squared: 0.8065
F-statistic:  3933 on 5 and 4711 DF,  p-value: < 2.2e-16
```

Tableau 7.17 Régression multiple : rang de monosémie (VD) avec maintien du rang de spécificité

La première colonne des valeurs estimées (*estimate*) montre que le rang de fréquence technique et le rang de spécificité ont un rapport de corrélation négative avec le rang de monosémie. Plus les mots sont fréquents dans le corpus technique et plus ils sont spécifiques, moins ils sont monosémiques. Cette observation corrobore la conclusion de l'analyse de régression simple pour le rang de spécificité. Le coefficient de la longueur indique également un rapport de corrélation négative : plus les mots sont longs, plus ils sont monosémiques. Finalement, nous observons un léger impact du nombre de classes lexicales : si un mot-clé appartient à plusieurs classes lexicales à la fois, il est plus hétérogène sémantiquement. Notons que l'appartenance à plusieurs classes lexicales pourrait s'interpréter comme un cas d'homonymie. La corrélation positive légèrement significative confirme donc l'hétérogénéité sémantique des homonymes, compte tenu du fait que notre mesure

de monosémie ne permet pas d'effectuer une distinction opérationnelle entre l'homonymie et la polysémie (Cf. chapitre 5). La dernière colonne de la valeur p montre que le rang de fréquence technique, le rang de spécificité et la longueur sont les facteurs les plus pertinents pour prédire le rang de monosémie des 4717 spécificités du corpus technique. Il est à noter que le rang de fréquence technique est la seule variable qui atteigne le plus haut degré possible de pertinence ($p < 2e^{-16}$).

Nos résultats confirment les observations formulées pour l'étude quantitative de la polysémie en langue générale (Oguy 1999). Oguy fait état d'une corrélation positive notamment entre la fréquence des mots et la polysémie d'une part et entre la structure morphologique simple et la longueur limitée des mots et la polysémie d'autre part. Les mots plus fréquents, plus courts et morphologiquement plus simples sont plus enclins à la polysémie. Il va sans dire que les mots les plus courts sont aussi les plus fréquents (Cf. la loi de Zipf) (Manning & Schütze 2002).

Pour la liste de 4717 spécificités, nous aimerions approfondir ces observations afin de fournir des réponses linguistiques plus appropriées au corpus technique. Il est clair que les spécificités les plus fréquentes du corpus technique, souvent à la fois les plus spécifiques, sont généralement les plus polysémiques, à quelques exceptions près. En plus, ce sont souvent les mots les plus courts et les plus simples morphologiquement. Comme nous avons signalé ci-dessus (Cf. chapitre 6), les mots les plus fréquents, tels que *machine* et *outil*, entrent très souvent dans la composition d'unités polylexicales (*machine à fraiser, machine à usiner, …*), ce qui pourrait en partie expliquer leur hétérogénéité sémantique ou polysémie. Il en va de même pour les mots les plus courts : ils se prêtent facilement à la composition de mots composés avec trait d'union ou d'unités polylexicales, d'où la corrélation positive avec les rangs de monosémie plutôt élevés. Rappelons que les unités polylexicales constituent une piste de recherche très intéressante que nous nous proposons d'explorer ultérieurement. Par ailleurs, le chapitre suivant consacré aux analyses de régression détaillées (Cf. chapitre 8), étudiera entre autres un sous-ensemble de mots composés, avec trait d'union et avec barre oblique, dont certains sont plutôt longs. Cette analyse permettra de jeter une lumière sur l'analyse des unités polylexicales et de vérifier si les conclusions formulées dans cette section se vérifient aussi pour un groupe de mots composés, catégorisés comme tels par Cordial.

- *Maintien du degré de spécificité*

La deuxième analyse de régression multiple ressemble beaucoup à l'analyse principale, à cette différence près que le rang de spécificité est supprimé en raison de son VIF trop élevé et que le degré de spécificité (log_LLR) est maintenu. Le rang de fréquence générale est également remplacé par l'écart des rangs de fréquence. Dans cette deuxième analyse de régression multiple pas à pas, les variables indépendantes

significatives (Cf. tableau 7.18) expliquent 80,68% de la variation du rang de monosémie[198]. Les variables indépendantes significatives sont très similaires aux variables indépendantes significatives du modèle principal, mais la corrélation négative du rang de spécificité est remplacée par la corrélation positive du degré de spécificité (log_LLR). Plus le degré de LLR est élevé (c'est-à-dire plus les mots sont spécifiques), plus ils sont polysémiques (rangs de monosémie près de 4700).

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)  4157.43062   67.14828  61.914  < 2e-16 ***
log_LLR       156.50174   19.90169   7.864 4.59e-15 ***
rang_v_freq1   -0.85304    0.01121 -76.101  < 2e-16 ***
long          -19.84211    2.73853  -7.246 5.01e-13 ***
nbr_claslex    69.30520   23.46212   2.954  0.00315 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 599.4 on 4712 degrees of freedom
Multiple R-Squared: 0.8069,     Adjusted R-squared: 0.8068
F-statistic:  4923 on 4 and 4712 DF,  p-value: < 2.2e-16
```

Tableau 7.18 Régression multiple : rang de monosémie (VD) avec maintien du degré de spécificité

7.2.2.2 Le rang de monosémie technique

- *Maintien du rang de spécificité*

Cette analyse de régression multiple étudie le pourcentage de variation expliquée R² et les corrélations, en prenant comme variable dépendante le rang de monosémie technique des 4717 spécificités. Le rang de spécificité est maintenu et le rang de fréquence générale est remplacé par l'écart des rangs de fréquence.

Les variables indépendantes significatives (Cf. tableau 7.19) expliquent 75,31% de la variation du rang de monosémie technique. Comme ce pourcentage est inférieur au pourcentage pour le rang de monosémie (80,65%), il confirme donc le pourcentage inférieur constaté pour le rang de monosémie technique dans l'analyse de régression simple (Cf. ci-dessus 7.1). Les variables indépendantes significatives sont les mêmes que celles pour le rang de monosémie. Les principales différences résident dans le seuil de significativité des variables et dans la corrélation positive du rang de spécificité.

[198] Ce pourcentage correspond à une différence de 0,03% par rapport au modèle principal du tableau 7.17.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4495.70159   49.34888  91.100  < 2e-16 ***
rang_v_spec     0.02836    0.01164   2.436  0.01488 *
rang_v_freq1   -0.90139    0.01265 -71.244  < 2e-16 ***
long          -21.60105    3.10100  -6.966 3.71e-12 ***
nbr_claslex    52.68923   26.54372   1.985  0.04720 *
freqabs1        0.06253    0.02413   2.592  0.00958 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 676.7 on 4711 degrees of freedom
Multiple R-Squared: 0.7534,     Adjusted R-squared: 0.7531
F-statistic:  2878 on 5 and 4711 DF,  p-value: < 2.2e-16
```

Tableau 7.19 Régression multiple : rang de monosémie technique (VD) avec maintien du rang de spécificité

Cette corrélation positive du rang de spécificité est très bizarre, à première vue, mais elle n'est pas cruciale dans le modèle, compte tenu de la pertinence plutôt faible (0,01) du rang de spécificité. Mais comment l'interpréter ? Les variables les plus significatives sont le rang de fréquence technique et la longueur. La faible pertinence du rang de spécificité indique que cette variable devra être considérée en tant que complément par rapport aux variables plus significatives. En effet, la variation qui reste inexpliquée, pourra être expliquée notamment par le rang de spécificité. Apparemment, cette fraction de la variation totale expliquée donne lieu à une corrélation positive entre le rang de spécificité et le rang de monosémie technique, ce qui permet d'expliquer la variation des spécificités à écarts importants. Signalons finalement la significativité de la fréquence absolue (technique) dans le modèle du rang de monosémie technique ainsi que la significativité très faible du nombre de classes lexicales. Celle-ci justifie par ailleurs le pourcentage de variation expliquée plus faible de 75,31%.

- *Maintien du degré de spécificité*

Nous procédons maintenant à la même analyse de régression multiple, mais pour le maintien du degré de spécificité (log_LLR). Les variables indépendantes significatives expliquent 75,3% de la variation du rang de monosémie technique (Cf. tableau 7.20), ce qui est parfaitement comparable au modèle précédent pour le rang de monosémie technique. Bien évidemment, ce pourcentage est inférieur au pourcentage obtenu pour le rang de monosémie (80,68%). Les variables indépendantes significatives sont relativement comparables à celles du modèle précédent pour le rang de monosémie technique. La significativité du nombre de classes lexicales et de la fréquence absolue dans le corpus technique est également plutôt faible.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.534e+03  4.919e+01  92.175  < 2e-16 ***
ecart_r_v_freq -2.374e-02  1.093e-02  -2.173   0.0298 *
rang_v_freq1   -8.835e-01  8.893e-03 -99.341  < 2e-16 ***
long           -2.185e+01  3.095e+00  -7.060 1.90e-12 ***
nbr_claslex     5.435e+01  2.656e+01   2.046   0.0408 *
freqabs1        5.680e-02  2.417e-02   2.350   0.0188 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 676.8 on 4711 degrees of freedom
Multiple R-Squared: 0.7533,     Adjusted R-squared: 0.753
F-statistic:  2877 on 5 and 4711 DF,  p-value: < 2.2e-16
```

Tableau 7.20 Régression multiple : rang de monosémie technique (VD) avec maintien du degré de spécificité

Toutefois, la différence la plus importante réside dans l'écart des rangs de fréquence (ecart_r_v_freq). En effet, même si ce modèle maintient le degré de spécificité (log_LLR), celui-ci ne figure plus dans le résultat final de la régression multiple pas à pas. Par contre, l'écart des rangs de fréquence est bel et bien significatif (valeur p comparable à celle du rang de spécificité dans le modèle précédent). L'écart des rangs de fréquence est une variable numérique avec des valeurs positives et négatives, moins faciles à interpréter en termes de proportionnalité.

Les visualisations de l'écart des rangs de fréquence, par rapport au rang de monosémie d'abord et par rapport au rang de monosémie technique ensuite (Cf. annexe 14 : figures A14.3 et A14.4), montrent effectivement que les mots les plus techniques (à écart positif et situés à droite) deviennent plus monosémiques, pour le rang de monosémie technique. En revanche, les mots les plus généraux (à écart négatif et situés à gauche) deviennent plus polysémiques, pour le rang de monosémie technique. Cette observation confirme donc que l'écart des rangs de fréquence joue un rôle complémentaire si on veut expliquer la variation du rang de monosémie technique, observé dans le modèle du tableau 7.20. L'écart des rangs de fréquence permet une subdivision à partir du degré plus ou moins technique ou plus ou moins général des mots, ce qui semble affecter le rang de monosémie technique, avec maintien du degré de spécificité.

7.2.2.3 Autres configurations des variables indépendantes

Afin d'étudier l'impact combiné de ces deux variables, nous procédons à des analyses de régression multiple supplémentaires, qui font intervenir en particulier la variable combinée (log_LLR et écart des rangs de fréquence).

- *Impact de la variable combinée (log_LLR et écart)*

La variable combinée est donc reprise dans le modèle de régression multiple pas à pas et le rang de fréquence générale est remplacé par l'écart des rangs de fréquence. Il est évident que dans la configuration avec maintien du rang de spécificité et suppression du log_LLR, il est impossible de procéder à l'analyse avec la variable combinée, parce qu'une de ses composantes (log_LLR) fait défaut. Dès lors, nous procédons uniquement à l'analyse de régression multiple avec la variable combinée pour la configuration sans rang de spécificité et avec maintien du degré de spécificité (log_LLR). Avant de passer à l'analyse proprement dite, signalons encore que l'analyse des VIF montre que l'écart des rangs de fréquence et la variable combinée sont intercorrélées. Après suppression, soit de l'écart soit de la variable combinée, l'autre variable se voit supprimée par l'analyse pas à pas, parce qu'elle n'est pas significative. Par conséquent, les résultats de cette analyse de régression multiple pas à pas sont identiques aux résultats précédents pour le maintien du degré de spécificité (log_LLR), tant pour le rang de monosémie (80,68%) que pour le rang de monosémie technique (75,3%). Les variables significatives sont donc identiques.

- *Impact de la variable catégorielle (classe lexicale)*

Finalement, les dernières analyses de régression multiple pas à pas font intervenir également la variable catégorielle de la classe lexicale. Comme nous l'avons évoqué ci-dessus, cette variable est exprimée par six valeurs différentes, dont une valeur de référence. Il s'ensuit que cinq valeurs sont reprises dans les résultats. Ces analyses sont conduites pour le rang de monosémie et le rang de monosémie technique, d'abord avec maintien du rang de spécificité et ensuite avec maintien du degré de spécificité. Les détails des résultats sont visualisés dans le document en annexe (Cf. annexe 14).

Globalement, les résultats sont similaires aux résultats des analyses principales pour le rang de monosémie (R^2 81,13%) et le rang de monosémie technique (R^2 75,73%) (Cf. 7.2.2.1 et 7.2.2.2). Cependant, nous tenons à insister sur une différence très importante concernant l'impact de la variable catégorielle de la classe lexicale. Pour le rang de monosémie, la classe lexicale 'adverbe' a un léger impact significatif (valeur p de 0,016 et de 0,025). De même, la classe lexicale 'nom' a un léger impact significatif pour le rang de monosémie (valeur p de 0,039) (maintien du degré de spécificité), ainsi que pour le rang de monosémie technique (valeur p de 0,045) (maintien du rang de spécificité). Les autres classes lexicales n'ont pas d'impact significatif sur les rangs de monosémie et de monosémie technique. Dans certains modèles de régression multiple, aucune classe lexicale n'est même significative.

Lorsque la classe lexicale du mot de base fait partie du modèle de régression multiple, généralement une seule valeur est significative pour cette variable

catégorielle, à savoir ‘adverbe’ ou ‘nom’. En plus, la pertinence de la classe lexicale significative se situe clairement dans la zone marginale du seuil de significativité, c’est-à-dire légèrement inférieur à 0,05. En conclusion, l’intégration de la variable catégorielle de la classe lexicale ne semble pas affecter de manière décisive les résultats des différentes analyses de régression multiple, ni en termes de R^2, ni en termes de variables significatives. Si l’on observe une significativité pour les valeurs de cette variable, la significativité est marginale et ne caractérise que deux classes lexicales.

Par conséquent, ces analyses de régression multiple qui intègrent la classe lexicale ne peuvent être les analyses principales. Or, elles sont intéressantes en guise de préparation aux analyses de régression détaillées. En effet, on pourrait se poser la question de savoir si l’impact léger des classes lexicales ‘adverbe’ et ‘nom’ se manifeste également dans les analyses détaillées par classe lexicale et si les résultats de celles-ci pour ces deux classes particulières se distinguent du reste des résultats (Cf. chapitre 8).

7.2.3 Conclusion de l’analyse de régression multiple

Les analyses de régression multiple (Cf. 7.2.2.1 et 7.2.2.2) pour le rang de monosémie et pour le rang de monosémie technique, tout comme les analyses supplémentaires qui font intervenir d’autres variables indépendantes (Cf. 7.2.2.3), permettent de formuler quelques conclusions générales pour nos 4717 spécificités.

En cas de maintien aussi bien du rang de spécificité que du degré de spécificité parmi les variables indépendantes du modèle de régression multiple, les variables indépendantes significatives expliquent environ 80% de la variation du rang de monosémie et 75% de la variation du rang de monosémie technique. La variable indépendante la plus significative est le rang de fréquence technique : plus les spécificités sont fréquentes dans le corpus technique, plus elles sont hétérogènes sémantiquement. Les autres variables indépendantes significatives sont le rang de spécificité ou respectivement le degré de spécificité, la longueur et finalement le nombre de classes lexicales. D’une part, pour le rang de monosémie, soit le rang de spécificité, soit le degré de spécificité se caractérisent par une corrélation négative : les mots les plus spécifiques sont les plus hétérogènes sémantiquement. D’autre part, pour le rang de monosémie technique, le rang de spécificité et l’écart des rangs de fréquence, jouent un rôle plutôt complémentaire par rapport aux autres variables indépendantes, en raison de leur significativité plus limitée.

Bien que les analyses de régression multiple supplémentaires intègrent des variables supplémentaires intéressantes à première vue, à savoir la variable combinée (log_LLR et écart des rangs de fréquence) et la variable catégorielle de la classe lexicale, ces nouvelles variables n’affectent pas de manière décisive les résultats des

analyses de régression multiple. Toutefois, certaines classes lexicales manifestent tout de même une faible pertinence qui constitue un indice intéressant pour procéder à des analyses de régression détaillées, entre autres par classe lexicale. Dans ces analyses, nous étudierons aussi le sous-ensemble des mots composés avec trait d'union ou avec barre oblique, comme nous l'avons annoncé ci-dessus.

Chapitre 8
Analyses de régression détaillées

Après la discussion des résultats des analyses de régression de base pour les 4717 spécificités (Cf. chapitre 7), nous présenterons dans ce chapitre les résultats des analyses de régression détaillées, c'est-à-dire pour divers sous-ensembles des 4717 spécificités et pour les spécificités des quatre sous-corpus du corpus technique (Cf. figure 3.1). La première partie de ce chapitre sera consacrée aux analyses de régression simple et multiple par classe lexicale (8.1). Il s'agit des sous-ensembles importants des substantifs et des adjectifs et des deux sous-ensembles plus restreints des verbes et des adverbes. Nous étudierons une fois de plus la corrélation entre le rang de monosémie (technique) et le rang de spécificité, pour un sous-ensemble déterminé des 4717 spécificités. Dans la deuxième partie (8.2), les analyses de régression seront conduites par sous-corpus (revues, fiches, normes, manuels), mais à partir de quatre nouvelles listes de spécificités. Nous terminerons ce chapitre par une conclusion globale (8.3). Notre objectif principal ici est de vérifier si la corrélation négative entre le rang de monosémie et le rang de spécificité observée lors des analyses de base se maintient, particulièrement dans le sous-corpus des normes.

8.1 ANALYSES DE RÉGRESSION PAR CLASSE LEXICALE

Les analyses de régression par classe lexicale sont conduites pour les classes lexicales des mots pleins : les substantifs, les adjectifs, les verbes et les adverbes. Pour certaines classes, nous procédons à des analyses plus détaillées, en particulier pour les substantifs déverbaux ainsi que pour les adverbes en *–ment*[199]. Les

[199] La classe lexicale des adverbes comprend essentiellement des adverbes en *–ment*, étant donné que les autres adverbes ont été intégrés dans la liste de mots grammaticaux. Toutefois, la classe lexicale comprend quelques adverbes (code Cordial 13) qui ne contiennent pas le morphème *–ment*, par exemple *plus* et *bien*, parce qu'ils portent en même temps le code d'une autre classe lexicale (*plus* : adverbe (8347 fois) et nom (30 fois)) (Cf. chapitre 3).

principales observations en termes de coefficients de corrélation et pourcentages de variation expliquée (ou de R²) (8.1.1) feront l'objet d'une analyse quantitative et linguistique (8.1.2).

Etant donné que les spécificités réparties par classe lexicale appartiennent à la liste de base des 4717 spécificités, elles se caractérisent par un double système de rangs (rangs de monosémie, de monosémie technique, de spécificité, de fréquence technique et de fréquence générale). D'une part, les spécificités par classe lexicale gardent leurs rangs de la liste des 4717 spécificités, ce qui explique pourquoi les rangs des 2923 substantifs varient entre 1 et 4717. Ce premier système de rangs (de 1 à 4717) vise principalement à situer les spécificités d'une certaine classe lexicale sur la visualisation de base, mais permet aussi d'étudier les corrélations et la variation expliquée par classe lexicale. D'autre part, les spécificités par classe lexicale se voient attribuer de nouveaux rangs à l'intérieur de la classe lexicale. Ainsi, les nouveaux rangs (de 1 à 2923 pour les substantifs), permettent d'analyser les corrélations et la variation expliquée à l'intérieur de la classe lexicale.

Les 4717 spécificités sont donc réparties en quatre sous-ensembles, en fonction de leur classe lexicale. Selon le cas, celle-ci peut être la classe à laquelle les mots appartiennent de manière exclusive ou non. Si un mot appartient à plusieurs classes lexicales, c'est la classe lexicale dominante qui a été retenue. Par exemple, le mot *mécanisme* appartient exclusivement à la classe lexicale des substantifs, tandis que le mot *mécanique* appartient à deux classes lexicales : il a été catégorisé 848 fois comme adjectif et 211 fois comme substantif. Par conséquent, pour *mécanique*, la classe lexicale des adjectifs a été retenue (Cf. chapitre 7). Le tableau 8.1 ci-dessous visualise le nombre de spécificités[200] par classe lexicale, ainsi que le pourcentage de spécificités par classe lexicale par rapport à la liste des 4717 spécificités.

	nombre	pourcentage
substantifs	2923	62%
adjectifs	1083	23%
verbes	541	11%
adverbes	141	3%

Tableau 8.1 Répartition des 4717 spécificités par classe lexicale

[200] La somme des 4 cases n'égale pas 4717 (mais 4688), parce que les spécificités avec la classe dominante « nprop » (27) ou « func » (2) ne font pas l'objet d'analyses de régression par classe lexicale (Cf. analyses par sous-catégorie : sigles). La répartition détaillée des 4717 spécificités, ainsi que la répartition de tous les lemmes du corpus technique et du corpus de référence sont comparées et visualisées en annexe (Cf. annexe 15).

8.1.1 Observations

Avant d'interpréter les données relatives aux classes lexicales, nous formulons un certain nombre d'observations importantes concernant les coefficients de corrélation (8.1.1.1), les résultats des analyses de régression (R^2) (8.1.1.2) et les variables significatives des analyses de régression multiple (8.1.1.3).

8.1.1.1 Coefficients de corrélation

Le tableau synoptique ci-dessous (Cf. tableau 8.2) montre la corrélation entre le rang de monosémie et le rang de spécificité par classe lexicale. Il indique, de même, la corrélation entre le rang de monosémie technique et le rang de spécificité. Dans les deux cas, cela vaut tant pour les rangs de base de 1 à 4717 que pour les nouveaux rangs. Les coefficients de corrélation Pearson sont partout négatifs et statistiquement significatifs, confirmant donc la corrélation négative entre le rang de monosémie (technique) et le rang de spécificité, que nous avons observée également pour les 4717 spécificités (Cf. chapitre 7) et rappelée en haut du tableau 8.2. Notons que la corrélation est plus faible pour le rang de monosémie technique que pour le rang de monosémie tout court. Les meilleures corrélations s'observent pour les substantifs : elles dépassent même les corrélations pour les 4717 spécificités. Les adverbes affichent les corrélations les plus faibles. Rappelons que les corrélations négatives signifient que les mots les plus spécifiques ne sont pas les plus monosémiques, au contraire.

	coefficient de corrélation Pearson : rang de monosémie (technique) ~ rang de spécificité	
mots 4717	rangs 1-4717	nouveaux rangs
mono	-0,71	
mono tech	-0,65	
adj 1083	rangs 1-4717	rangs 1-1083
mono	-0,69	-0,70
mono tech	-0,62	-0,63
adv 141	rangs 1-4717	rangs 1-141
mono	-0,60	-0,62
mono tech	-0,53	-0,55
nom 2923	rangs 1-4717	rangs 1-2923
mono	-0,74	-0,74
mono tech	-0,68	-0,68
verbe 541	rangs 1-4717	rangs 1-541
mono	-0,66	-0,67
mono tech	-0,59	-0,60

Tableau 8.2 Corrélations par classe lexicale

8.1.1.2 Résultats des analyses de régression : R²

Le tableau comparatif 8.3 visualise les résultats des analyses de régression simple et multiple pour les quatre classes lexicales, aussi bien pour la variable dépendante du rang de monosémie que pour celle du rang de monosémie technique. Les analyses de régression simple sont conduites pour les rangs de 1 à 4717 et pour les nouveaux rangs. Les analyses de régression multiple ont été effectuées à l'intérieur de la classe lexicale, à partir des nouveaux rangs, et sont soumises, pour éviter le problème de multicolinéarité, à la vérification préalable des VIF des variables indépendantes[201]. Les résultats des analyses de base (mots 4717) sont repris à titre d'information.

	simple R²		**multiple R²**	VI → rvfq2 remplacé par écart ;
	r 0-4717	nouv. rangs	nouv. rangs	log:ecart (si log_LLR aussi inclus)
mots 4717				
mono	51,57% hé[202]		80,65%	rvfq1 ; rvspec ; long ; *nbr_claslex*
mono tech	42,74% hé		75,31%	rvfq1 ; long ; *fqabs1 ; rvspec* ; *nbr_claslex*
adj 1083				
mono	48,76% hé	49,48% hé	77,18%	rvfq1 ; *écart ; nbr_claslex*
mono tech	39,08% hé	39,87% hé	71,12%	rvfq1 ; *nbr_claslex ; fqabs1*
adv 141				
mono	36,61% hé	38,31% hé	70,13%	rvfq1 ; *log_LLR* ; long
mono tech	27,98% hé	30,55% hé	66,52%	rvfq1 ; long
nom 2923				
mono	55,77% hé	54,75% hé	81,95%	rvfq1 ; long ; rvspec ; *nbr_claslex*
mono tech	47,48% hé	46,37% hé	76,12%	rvfq1 ; long ; écart ; *nbr_claslex*
verbe 541				
mono	43,50% hé	45,20% hé	82,30%	rvfq1 ; rvspec ; long
mono tech	34,96% hé	36,29% hé	78,80%	rvfq1 ; *fqabs2*

Tableau 8.3 Résultats des analyses de régression par classe lexicale

[201] Pour les classes lexicales des adjectifs, des substantifs et des verbes, le log_LLR est supprimé et le rang de fréquence générale (rvfq2) est remplacé par l'écart des rangs de fréquence, comme dans l'analyse multiple de base. Par contre, pour les adverbes, la vérification des VIF mène à la suppression de la fréquence absolue dans le corpus technique (fqabs1) et du rang de spécificité (rvspec), ce qui permet de garder le log_LLR et d'inclure la variable combinée (log_LLR et écart), le rang de fréquence générale étant remplacé par l'écart des rangs de fréquence (Cf. annexe 15).

[202] Abréviations : hé = hétéroscédasticité ; ho = homoscédasticité.

Selon le test de Goldfeld-Quandt (gqtest), les quatre sous-ensembles de spécificités réparties par classe lexicale se caractérisent par l'hétéroscédasticité, tout comme les 4717 spécificités. Pour les analyses détaillées, nous ne procédons pas aux solutions techniques, ni aux solutions de répartition et d'exclusion, auxquelles nous avons eu recours pour les analyses de base (Cf. chapitre 7). Nous aimerions plutôt vérifier si la conclusion générale que nous avons formulée pour les analyses de base, se confirme dans les analyses détaillées. Nous nous demandons par ailleurs si l'explication du problème de l'hétéroscédasticité s'applique aussi aux analyses par classe lexicale : est-ce que le fait que les mots les plus fréquents du corpus général échappent au pouvoir prédictif du modèle de régression simple se reproduit ici ? Les visualisations des régressions simples par classe lexicale en annexe (Cf. annexe 15 : figures A15.4 à A15.19) montrent en effet clairement des mots à résidus importants, en haut à droite, qui ne suivent pas la tendance globale de corrélation négative.

La comparaison des pourcentages de variation expliquée R^2 (Cf. tableau 8.3) indique les pourcentage de R^2 les plus élevés pour les substantifs (± 55% rang de monosémie et ± 47% rang de monosémie technique) et les pourcentages de R^2 les moins élevés pour les adverbes (± 37-38% et ± 28-30%). Les résultats des substantifs dépassent même les résultats des analyses de régression simple de base (51% et 42%). Par conséquent, les substantifs se prêtent légèrement mieux à la tendance de corrélation négative, bien qu'elle ne soit pas tout à fait linéaire. Les résultats des analyses de régression multiple confirment ces tendances en termes de R^2, bien que les pourcentages de R^2 des verbes (82% et 78%) dépassent légèrement ceux des substantifs (82% et 76%) et ceux des 4717 spécificités (80% et 75%). Les adverbes ont les pourcentages de R^2 les plus faibles.

8.1.1.3 Analyses de régression multiple : variables significatives

Les variables indépendantes qui sont significatives pour les analyses de régression multiple sont visualisées dans la dernière colonne du tableau ci-dessus (Cf. tableau 8.3). Celles qui se caractérisent par une corrélation positive avec la variable dépendante (rang de monosémie ou rang de monosémie technique) sont indiquées en italique. Les autres ont donc une corrélation négative avec la variable dépendante, en particulier le rang de fréquence technique (rvfq1). Les variables indépendantes sont classées par ordre décroissant de significativité (valeur p).

Pour les quatre classes lexicales, tant pour le rang de monosémie que pour le rang de monosémie technique, le rang de fréquence technique est la variable indépendante la plus significative et se caractérise par une corrélation négative avec la variable dépendante. Il s'ensuit que, dans un modèle qui inclut toutes les variables indépendantes significatives, le rang de fréquence technique explique le mieux la variation du rang de monosémie ou du rang de monosémie technique. En plus, les

spécificités réparties par classe lexicale les plus fréquentes dans le corpus technique sont les moins monosémiques et, dès lors, les plus hétérogènes sémantiquement. Inversement, les spécificités les moins fréquentes dans le corpus technique sont les plus monosémiques, ce qui confirme les observations que nous avons faites à partir des analyses de base (Cf. mots 4717).

Les observations concernant la longueur et le nombre de classes lexicales se voient confirmées également, même si ce n'est que pour certaines classes lexicales. En effet, la longueur, qui n'est pas exprimée en termes de rangs mais en nombre de caractères, se caractérise par une corrélation négative avec le rang de monosémie (technique). Les adverbes et les substantifs les plus longs (comprenant le plus de caractères), telles que *perpendiculairement* et *affûteuse-rectifieuse* sont les plus monosémiques (rangs de monosémie moins élevés ou près de 1) alors que les adverbes et les substantifs les moins longs (*plus, bien* et *axe, air*) sont les moins monosémiques (rangs de monosémie plus élevés). Il est à noter cependant que la longueur n'est pas significative pour la classe lexicale des adjectifs. Ajoutons à cela que la variable du nombre de classes lexicales s'avère significative uniquement pour les adjectifs et pour les substantifs. En effet, ce sont principalement ces deux classes lexicales qui sont impliquées dans les étiquettes à plusieurs classes lexicales (2 ou 3 ou 4) (Cf. annexe 13). La corrélation positive est confirmée : les adjectifs et les substantifs qui appartiennent en même temps à une autre classe lexicale (respectivement celle des substantifs et des adjectifs) ont des rangs de monosémie plus élevés et sont dès lors plus hétérogènes sémantiquement. En l'occurrence, ils sont homonymiques, par exemple *technique, automatique, mécanique, manuel, standard*.

Finalement, les variables indépendantes qui correspondent au rang de spécificité et au degré de spécificité expliquent également en partie la variation de la variable dépendante. Le rang de spécificité est significatif pour les substantifs et pour les verbes, mais uniquement pour la variable dépendante du rang de monosémie. Pour le rang de monosémie technique des substantifs, c'est l'écart des rangs de fréquence qui est significatif. Ce dernier représente la différence (ou l'écart) entre les rangs de fréquence dans le corpus technique et dans le corpus général et indique la technicité du mot en question. La variable de l'écart est également significative pour les adjectifs, pour le rang de monosémie. Comme le degré de spécificité (ou le log_LLR) n'a pas été supprimé pour la classe lexicale des adverbes, il s'avère significatif pour le rang de monosémie : les adverbes les plus spécifiques (ayant la valeur de log_LLR la plus élevée) sont les moins monosémiques, par exemple *également, entièrement, généralement, directement, facilement*.

8.1.2. Interprétations

Comme nous l'avons évoqué ci-dessus (Cf. 8.1.1), le rang de spécificité se caractérise à travers les différentes classes lexicales par une corrélation négative avec le rang de monosémie et avec le rang de monosémie technique : en d'autres mots, les spécificités les plus spécifiques sont donc les moins monosémiques. Ainsi, les observations par classe lexicale confirment nos observations antérieures relatives aux analyses de base et remettent en question, une fois de plus, la thèse des monosémistes. Cela est vrai avant tout pour la classe lexicale des substantifs. Rappelons à ce sujet que les textes techniques se distinguent des textes « de langue générale », pour autant que cette dichotomie soit légitime, par une surabondance de substantifs (Kocourek 1991a). En effet, dans notre liste de 4717 spécificités, les substantifs sont bien représentés, constituant même la majorité (62%) des 4717 spécificités. Si la thèse des monosémistes qui prône la monosémie dans les textes techniques se vérifiait, elle serait d'autant plus vraie pour les unités les plus spécifiques (les 4717 spécificités) et pour les unités de la classe lexicale la plus représentée (les substantifs). Or, les 4717 mots les plus spécifiques du corpus technique se caractérisent par une corrélation négative entre le rang de spécificité et le rang de monosémie (coefficient de corrélation Pearson de -0,71 et R^2 de 51,57%) (Cf. chapitre 7), puisque les mots les plus spécifiques sont les plus hétérogènes sémantiquement. Dès lors, la thèse des monosémistes est infirmée pour les mots les plus spécifiques du corpus technique (Cf. conclusion chapitre 7). Cette corrélation négative est même plus forte encore pour les 2923 substantifs (coefficient de corrélation Pearson de -0,74 et R^2 de 55,77%), ce qui ébranle définitivement la thèse des monosémistes.

Rappelons que pendant les analyses de base du chapitre précédent, nous avions isolé un sous-ensemble de 1507 spécificités fréquentes dans le corpus de langue générale, qui étaient responsables du problème de l'hétéroscédasticité. Celles-ci entraînaient un effet perturbateur pour l'ensemble des 4717 spécificités, dans la mesure où elles échappaient à la tendance de corrélation négative entre le rang de spécificité et le rang de monosémie. Dans le but de vérifier si les spécificités les plus générales par classe lexicale ont le même effet perturbateur et afin d'interpréter les résultats du tableau 8.3, nous procéderons à une explication quantitative (8.1.2.1) et à une explication linguistique (8.1.2.2), ainsi qu'à des mises au point à partir de plusieurs sous-catégories de spécificités (8.1.2.3).

8.1.2.1 Explication quantitative

Les analyses de régression simple du tableau 8.3 ci-dessus affichent les meilleurs pourcentages de variation expliquée (R^2) pour les substantifs (± 55% rang de monosémie et ± 47% rang de monosémie technique). Les pourcentages les plus faibles s'observent pour les adverbes (37% et 28%), et dans une moindre mesure,

pour les verbes (43% et 35%). Ces pourcentages de R² plus faibles soulèvent bien sûr la question de savoir si les adverbes et les verbes comprennent plus de spécificités (plus) générales, puisqu'on sait que les spécificités les plus générales sont susceptibles d'entraîner un effet perturbateur.

Le tableau ci-dessous (Cf. tableau 8.4) visualise la répartition des 4717 spécificités par classe lexicale (Cf. tableau 8.1), ainsi que celle des 1507 spécificités générales à effet perturbateur. Les verbes et les adverbes sont effectivement mieux représentés dans ce sous-ensemble (respectivement 19% et 5%) que dans la liste entière (11% et 3%), où les substantifs sont plus nombreux (62%). Les classes lexicales des verbes et des adverbes comprennent donc relativement plus de « spécificités à effet perturbateur ». Cette comparaison permet d'expliquer, non seulement les meilleurs pourcentages de variation expliquée R² des substantifs (proportionnellement moins bien représentés dans le sous-ensemble des 1507 spécificités), mais également les pourcentages de R² plus faibles des verbes et des adverbes.

	nombre (4717)	% (4717)	nombre (1507)	% (1507)
substantifs	2923	**62%**	770	51%
adjectifs	1083	23%	382	25%
verbes	541	11%	286	**19%**
adverbes	141	3%	68	**5%**

Tableau 8.4 Répartition des 4717 et des 1507 spécificités par classe lexicale

La moyenne du rang de fréquence générale (rangs de 1 à 4717) ne peut que confirmer cette explication quantitative. En effet, les substantifs sont globalement les moins fréquents dans le corpus général, avec une moyenne du rang de fréquence générale de 1216[203]. Par contre, les adverbes (moyenne de 69) et, tout de suite après, les verbes (moyenne de 266) sont globalement les plus fréquents dans le corpus général. Pour les nouveaux rangs, la moyenne de la fréquence absolue dans le corpus général confirme d'une part la fréquence générale plus élevée des adverbes (moyenne de fréquence absolue de 942) et des verbes (moyenne de 484) et d'autre part la fréquence générale plus limitée des substantifs (moyenne de 216).

Comme la fréquence moyenne des substantifs dans le corpus général est moins élevée que celle des autres classes lexicales, ils affichent de meilleurs pourcentages de R² ainsi que de meilleures corrélations négatives. Il s'ensuit que le rang de spécificité des substantifs permet d'expliquer ou de prédire même leur rang de monosémie et leur rang de monosémie technique. Les adverbes sont globalement les

[203] Rappelons que des rangs plus élevés correspondent à une fréquence générale plus faible.

plus fréquents dans le corpus général, ce qui se traduit par des pourcentages de R^2 plus faibles : les adverbes suivent moins bien la tendance générale de corrélation négative. En raison du caractère plus général des adverbes, leur rang de monosémie et leur rang de monosémie technique sont moins faciles à expliquer ou à prédire à partir de leur rang de spécificité.

8.1.2.2 Explication linguistique

Les résultats des analyses de régression par classe lexicale se prêtent également à une explication essentiellement linguistique. En effet, l'analyse des cooccurrences, visant à déterminer le degré de monosémie et, dès lors, le rang de monosémie des spécificités, est tributaire de leurs caractéristiques syntaxiques. Les effets observés dans le tableau 8.3 et les différences en termes de variation expliquée (R^2) sont effectivement liés aux caractéristiques syntaxiques des spécificités et plus particulièrement à leurs propriétés collocationnelles.

Selon la classe lexicale, les spécificités se comportent différemment pour ce qui est des collocations et des cooccurrences. En effet, le mécanisme collocationnel des adverbes est moins puissant que celui des substantifs ou des adjectifs, par exemple. Les substantifs sont désambiguïsés par des adjectifs qualificatifs, par des déterminants et par des verbes, avec lesquels ils ont des relations collocationnelles très fortes. Par conséquent, les substantifs ont relativement plus de cooccurrents stables et statistiquement très significatifs. Les adjectifs et les verbes en particulier, forment souvent de vraies collocations avec les substantifs, par exemple *avance technologique* (« progression »), *augmenter l'avance (d'un outil)* (« la vitesse »), *usiner une pièce*. De même, les adjectifs sont principalement désambiguïsés par les substantifs qu'ils modifient ou qu'ils qualifient et qui constituent également des cooccurrents stables et statistiquement très significatifs, par exemple *outil rotatif* (« qui tourne autour d'un axe ») versus *table rotative* (« que l'on fait tourner »). Il en va de même pour les verbes, qui sont désambiguïsés par leurs arguments (sujet, COD, COI), généralement des substantifs, par exemple *usiner des trous* (« tarauder ») versus *usiner des pièces* (« fraiser, rectifier »). Par contre, le mécanisme désambiguïsateur et collocationnel des adverbes est généralement moins clair : l'applicabilité de l'analyse des cooccurrences est donc plus restreinte pour la classe lexicale des adverbes, dans la mesure où ceux-ci ont peu de cooccurrents stables ou statistiquement très significatifs.

Par conséquent, le pourcentage limité de variation expliquée (R^2) des adverbes, c'est-à-dire le fait que la variation du rang de spécificité des adverbes ne permet pas de rendre compte de manière satisfaisante de la variation quant au rang de monosémie, pourrait s'expliquer par l'applicabilité plus restreinte de notre mesure de monosémie, basée sur l'analyse des cooccurrences. Pour les verbes, la raison du

pourcentage plutôt limité de variation expliquée (R^2) pourrait résider dans leur position intermédiaire entre les adverbes, d'une part, et les substantifs et adjectifs, de l'autre. Les verbes sont désambiguïsés par leurs arguments, donc par les substantifs qu'ils sélectionnent suivant le principe des restrictions de sélection. Si certains verbes sélectionnent clairement une petite série de substantifs bien déterminés (sélection restreinte), il s'avère que pour d'autres verbes, plus généraux, la sélection d'arguments est moins contraignante ou moins restreinte. Ainsi, des verbes plutôt généraux[204] et fréquents dans le corpus général, tels que *permettre, présenter* et *proposer*, sélectionnent des arguments sémantiquement très différents les uns des autres, ce qui explique leur hétérogénéité sémantique.

Il est clair que cette explication se heurte à la frontière technique de l'analyse des cooccurrences. Elle requiert donc une analyse des cooccurrences plus fine, c'est-à-dire une analyse des cooccurrences « enrichie », qui intègre également les caractéristiques syntaxiques[205] des mots de base (ou spécificités) et de leurs cooccurrents.

8.1.2.3 Conclusion et mises au point

Pour conclure, nous commenterons la cohérence des résultats tout en les complétant par une série d'analyses détaillées pour les sous-catégories des substantifs déverbaux et des adverbes en *–ment*, des sigles et des mots composés avec trait d'union ou avec barre oblique.

Globalement, rappelons-le, la classe lexicale des substantifs illustre le mieux la corrélation négative entre le rang de monosémie (technique) et le rang de spécificité. Elle confirme de ce fait le pouvoir explicatif et prédictif du rang de spécificité, dans la régression simple, et de toutes les variables indépendantes significatives, dans la régression multiple. Il se trouve que les analyses de la sous-catégorie des substantifs déverbaux (en *–ion*, en *–age* et en *–ment*) non seulement confirment ces résultats mais affichent même de meilleurs résultats. Pour les analyses de régression simple, le pourcentage de R^2 s'élève à 58-59% (rang de monosémie) et à 52-53% (rang de

---

[204] Les verbes les plus polysémiques du corpus technique sont en même temps les plus fréquents dans le corpus général : ils ont des nouveaux rangs de fréquence générale (de 1 à 541) inférieurs à 100 et des rangs de fréquence générale de base (de 1 à 4717) inférieurs à 500. Citons les verbes les plus fréquents du corpus général qui sont en outre les plus hétérogènes sémantiquement : *présenter, développer, proposer, permettre, assurer, comprendre, concerner, comporter, réaliser, intégrer, utiliser, prévoir, mesurer, destiner*.

[205] Le code Cordial indique la classe lexicale ou donne des informations supplémentaires sur le genre, le temps du verbe, la personne, etc.

monosémie technique) ; pour l'analyse de régression multiple, il s'élève à 82% et à 79% respectivement[206]. Pour ce qui est de l'aspect quantitatif, il convient de signaler que les substantifs déverbaux sont en moyenne moins fréquents dans le corpus général (moyenne de fréquence absolue dans le corpus général de 166) que les substantifs pris ensemble (moyenne de 216). Cette observation confirme donc la conclusion générale formulée précédemment pour les analyses de base : un sous-ensemble de spécificités qui comprend moins de spécificités fréquentes dans le corpus général, corrobore mieux le pouvoir explicatif des modèles de régression simple et multiple.

Par ailleurs, la classe lexicale des adverbes se prête moins bien à la corrélation négative entre le rang de monosémie (technique) et le rang de spécificité, non seulement parce que cette classe lexicale comprend en moyenne le plus de spécificités « à effet perturbateur », c'est-à-dire les plus fréquentes dans le corpus général, mais aussi en raison des propriétés syntaxiques et collocationnelles des adverbes. La suppression des adverbes qui ne sont pas en *–ment* (*plus, bien,* …), mais qui appartiennent à plusieurs classes lexicales à la fois, permet de supprimer les adverbes les plus fréquents dans le corpus général[207]. On aboutit par là même à l'homoscédasticité pour la sous-catégorie des adverbes en *–ment*. D'ailleurs, dans cette sous-catégorie, les pourcentages de variation expliquée sont plus élevés (39-42% pour le rang de monosémie et 30-34% pour le rang de monosémie technique) que dans la classe lexicale générale des adverbes (37% et 28%).

Pour la classe lexicale des verbes, la comparaison des différentes analyses de régression simple et multiple par classe lexicale met en évidence un manque de cohérence. En effet, les pourcentages de variation expliquée R^2 des analyses de régression simple ne sont pas très élevés, mais ceux des analyses multiples sont les meilleurs de toutes les classes lexicales. Signalons à cet effet le pouvoir explicatif important de la fréquence absolue des verbes dans le corpus général pour les analyses de régression multiple.

Après les analyses détaillées pour les sous-catégories des substantifs déverbaux et des adverbes en *–ment*, nous avons aussi procédé à des analyses détaillées pour d'autres sous-catégories, comprenant majoritairement des substantifs, mais aussi des

[206] Les détails des sous-catégories sont visualisés en annexe (Cf. annexe 15 : tableau A15.2).

[207] Pour les 141 adverbes, la moyenne de la fréquence absolue dans le corpus général est de 942, tandis que celle des 131 adverbes en *–ment* n'est que de 283, ce qui est une différence considérable.

spécificités d'autres classes lexicales. Il s'agit de la sous-catégorie des sigles[208] (spécificités à une, deux ou trois lettres) et de la sous-catégorie des mots composés[209], c'est-à-dire des spécificités avec trait d'union (-) ou avec barre oblique (/), catégorisées par Cordial comme une seule unité lexicale, même si elles se rapprochent parfois des unités polylexicales, par exemple *t/min, m/min*. Les détails des analyses de régression pour ces deux sous-catégories sont visualisés dans le tableau comparatif A15.1 en annexe (Cf. annexe 15).

La sous-catégorie des sigles se caractérise par le pourcentage de R^2 le plus élevé pour la régression multiple. La variation du rang de monosémie comme variable dépendante est expliquée principalement par le rang de fréquence technique, mais aussi par le rang de spécificité, puisqu'elle affiche une corrélation positive avec le rang de monosémie et s'avère donc complémentaire au rang de fréquence technique. En effet, quelques spécificités sont peu spécifiques et polysémiques (*non, air, eau*), d'autres spécificités sont plus spécifiques et moins polysémiques (*mm, t, z*), ce qui explique la corrélation positive complémentaire de la variable indépendante du rang de spécificité dans le modèle de régression multiple (Cf. visualisations des sous-catégories dans l'annexe 15).

La sous-catégorie des mots composés avec trait d'union ou avec barre oblique comprend presque seulement des spécificités qui sont absentes du corpus de référence de langue générale, à quelques exceptions près (*sous-traitance, technico-commercial, technico-économique, pick-up*). Il s'ensuit que la moyenne de la fréquence absolue dans le corpus général est particulièrement faible (0,2). En plus,

---

[208] La sous-catégorie des sigles (194) comprend 153 substantifs (78%), 16 adjectifs (*usé, lié, sec, …*), 5 adverbes (*non, dur, …*) et 20 noms propres qui entrent aussi dans la classe lexicale des substantifs (*Fao, Cao, …*). Le groupe des sigles comprend 3 items à quatre lettres (*Cfao, Gpao* et *Nbre*). Si nous avons retenu ces exceptions au principe des trois lettres, c'est pour des raisons évidentes de cohérence : ces notions constituent en effet des concepts-clés du domaine. Ainsi, nous avons retenu *Cao* et il aurait été peu logique de ne pas inclure également *Cfao*.

Il est à noter que la sous-catégorie des sigles comprend des mots à trois lettres qui ne sont pas du tout des abréviations ou des initiales, tels que *non, cas, air, eau, vue, sol, dur, clé, gaz, col, jet*. Le critère d'appartenance à la sous-catégorie est le critère objectif et quantitatif de la longueur de la spécificité (longueur de 1 à 3 caractères). Nous sommes tout à fait consciente de l'hétérogénéité de la sous-catégorie « sigles » et espérons remédier à cette lacune dans des travaux ultérieurs.

[209] La sous-catégorie des mots composés (429) se constitue de 368 substantifs (85%), de 54 adjectifs (*ultra-fin, ultra-rapide, élastico-plastique, …*), d'un adverbe (*entre-temps*), de 4 verbes (*sous-traiter, sous-estimer, …*) et de 2 noms propres (*L/min, …*).

cette sous-catégorie se caractérise par l'homoscédasticité (Cf. annexe 15) et dès lors par une visualisation de forme différente de celle des 4717 spécificités de base. La sous-catégorie des mots composés se caractérise aussi par des pourcentages de R^2 très élevés (61% et 52%) pour les analyses de régression simple (et pour les rangs de 1 à 4717). La corrélation entre le rang de spécificité et le rang de monosémie est négative, ce qui veut dire que, même pour les spécificités absentes du corpus général et très typiques de la langue spécialisée, les mots composés les plus spécifiques ne sont toujours pas les plus monosémiques, au contraire.

Toutefois, pour les nouveaux rangs (de 1 à 429), donc à l'intérieur de la sous-catégorie des mots composés, on observe une chute importante des pourcentages de variation expliquée R^2 (61 → 47% rang de monosémie et 52 → 38% rang de monosémie technique)[210]. Bien que la corrélation soit négative et statistiquement significative et donc fiable, on constate que, pour les nouveaux rangs de monosémie et de spécificité, les mots composés sont plus dispersés sur la visualisation (Cf. annexe 15). Leur nouveau rang de spécificité (de 1 à 429) permet moins bien de prédire leur rang de monosémie ou de monosémie technique. En effet, pour la même valeur du rang de spécificité, les valeurs du rang de monosémie sont très hétérogènes et dispersées.

La sous-catégorie des mots composés avec trait d'union ou avec barre oblique est apparentée à celle des unités polylexicales, en raison de leur caractère composé qui facilite d'ailleurs une certaine désambiguïsation[211]. Par conséquent, nous pourrions déjà avancer l'hypothèse que les unités polylexicales se prêteront moins bien à une corrélation négative entre le rang de monosémie et le rang de spécificité. Des recherches futures permettront de le vérifier, ou non, fondées sur de nouvelles analyses statistiques de régression. A ce sujet, nous envisageons de dissocier les deux composants des spécificités de cette sous-catégorie et de considérer le premier composant comme « mot de base » et le deuxième comme cooccurrent, dont les c seront considérés comme cc du mot de base. Cette dissociation nous permettra de vérifier l'effet de notre formule de monosémie pour les mots composés et de faire un premier pas vers l'analyse des unités polylexicales, qui nécessiteront de toutes façons une adaptation de la formule de monosémie.

---

[210] Les pourcentages de R^2 des analyses de régression multiple, effectuées pour les nouveaux rangs, sont aussi très faibles (47% rang de monosémie et 39% rang de monosémie technique).

[211] Le deuxième composant (par exemple dans *t/min*) a un effet désambiguïsateur pour le premier composant.

8.1.2.4 Interprétation de l'hétérogénéité sémantique

Finalement, nous revenons sur la question de l'hétérogénéité sémantique formulée en termes d'homonymie, de polysémie et de vague[212]. Nous proposons un certain nombre d'hypothèses linguistiques, différenciées en fonction de la classe lexicale des spécificités.

D'une part, pour la classe lexicale des substantifs, on pourrait avancer l'hypothèse que l'hétérogénéité sémantique correspond surtout à de la polysémie, c'est-à-dire à la présence de plusieurs sens apparentés sémantiquement. Les sens de certains substantifs déverbaux, tels que *filetage, fabrication*, hétérogènes sémantiquement selon notre mesure de monosémie, se caractérisent effectivement par un rapport métonymique (action – résultat).

D'autre part, pour la classe lexicale des adverbes, l'hétérogénéité sémantique pourrait se traduire par l'homonymie et par le vague. Les adverbes qui ne se terminent pas par *–ment*, tels que *plus, non, bien*, et qui appartiennent à plusieurs classes lexicales sont hétérogènes sémantiquement : ils pourraient dès lors être qualifiés d'homonymes. Certains adverbes en *–ment*, tels que *simplement, seulement, uniquement, également*, se caractérisent par un double emploi, adverbial et conjonctif, bien que l'emploi conjonctif soit marginal (Cf. annexe 5 : liste de mots grammaticaux). Les adverbes peu fréquents ou absents du corpus général et dérivés d'adjectifs techniques, par exemple *hydrauliquement* et *axialement,* sont plus spécifiques et plus homogènes sémantiquement. Ces différents types d'adverbes expliquent également les pourcentages de R^2 plutôt faibles, autant pour les adverbes en *–ment* (131) que pour l'ensemble des adverbes (141). Ces pourcentages sont en plus étroitement liés aux caractéristiques syntaxiques et collocationnelles des adverbes.

Cependant, pour vérifier les hypothèses que nous venons d'esquisser quant à l'hétérogénéité sémantique, des recherches statistiques multivariées supplémentaires s'imposent. Une analyse de regroupement (*cluster analysis*) permettra de regrouper les cooccurrents d'un mot de base ou spécificité, à partir des cc qu'ils partagent.

---

[212] Rappelons que le vague est un phénomène d'indétermination du référent.

8.2 ANALYSES DE RÉGRESSION PAR SOUS-CORPUS

Dans le but d'affiner les conclusions de notre étude, nous procédons aussi à des analyses de régression détaillées par sous-corpus. En effet, le corpus technique est constitué de quatre sous-corpus, qui manifestent différents niveaux de normalisation et de vulgarisation. Nous commençons par comparer leurs caractéristiques quantitatives (Cf. tableau 8.5 ci-dessous).

Etant donné que le sous-corpus des revues est deux fois plus vaste que les autres sous-corpus, nous préférons établir la comparaison à partir d'un échantillon aléatoire des revues de taille comparable[213]. Il est à noter que les formes graphiques et lemmes indiqués dans le tableau ci-dessous ne comprennent pas de signes de ponctuation[214], ni au niveau des *types*, ni au niveau des *tokens*. Comme nous l'avons évoqué ci-dessus (Cf. chapitre 3 : tableaux 3.8 et 3.9), le *Type-Token Ratio* ou TTR[215] (5), permet de mesurer la richesse ou la diversité lexicale d'un sous-corpus. Plus le TTR d'un sous-corpus est élevé, plus il contient de formes différentes ou de lemmes différents.

La comparaison des TTR des formes graphiques et des lemmes indique que les normes et les manuels ont des TTR similaires (4,1 et 2,6), inférieurs aux TTR des revues (5,8 et 3,6) et des fiches (6,7 et 4,2). Le sous-corpus des fiches est le plus diversifié lexicalement, probablement en raison des particularités stylistiques des fiches : signalons l'absence de texte suivi et l'énumération de caractéristiques techniques. Cette diversité lexicale s'exprime également par le rapport inverse du TTR, à savoir le *Token-Type Ratio* (7) et (8), qui indique la récurrence ou la répétition des formes graphiques ou des lemmes. Dans les normes et les manuels, la fréquence moyenne des formes graphiques est de 23, celle des lemmes de 38. Ces deux sous-corpus prescriptifs se caractérisent par une récurrence plus importante des formes graphiques et des lemmes, ainsi que par une homogénéité thématique plus importante.

---

[213] Dans un corpus plus long, les mots ont plus de chances d'être répétés, ce qui se traduit généralement par un TTR plus faible (Cf. tableau 8.5).

[214] Par rapport au total des signes (mots et signes de ponctuation), les signes de ponctuation représentent à peu près 9% dans les sous-corpus : 8,8% et 8,75% (revues et échantillon revues), 9,19% (normes), 9,09% (manuels). Toutefois, dans les fiches, les signes de ponctuation représentent 11,2%, confirmant la particularité typographique et stylistique des fiches (Cf. Chapitre 3 : 10,7% de signes de ponctuation dans le corpus technique et 8,6% dans le corpus de référence de langue générale).

[215] (Nombre de formes graphiques différentes *100) / nombre total de formes graphiques.

	revues	fiches	normes	manuels	revues échantillon
(1) Nombre total de formes graphiques (*tokens*)	790.680	296.650	286.139	378.331	304.977
(2) Nombre de formes graphiques différentes (*types*)	30.298	19.995	12.003	15.814	17.829
(3) Nombre total de lemmes (*tokens*)	790.680	296.650	286.139	378.331	304.977
(4) Nombre de lemmes différents (*types*)	18.449	12.658	7.441	9.927	11.055
(5) TTR formes graphiques	3,8318915	**6,7402663**	***4,1948144***	***4,1799377***	**5,8460146**
(6) TTR lemmes	2,333308	**4,2669813**	***2,6004844***	***2,6238928***	**3,6248635**
(7) *Token-Type Ratio* : formes graph.	26,096772	14,836209	23,838957	23,923802	17,105671
(8) *Token-Type Ratio* : lemmes	42,857607	23,435772	38,454374	38,111313	27,587246
(9) *Types* formes graphiques / lemmes	1,642257	1,5796334	1,6130896	1,5930291	1,6127544

Tableau 8.5 Lemmes et formes graphiques par sous-corpus

Les analyses de régression détaillées par sous-corpus ne seront pas conduites pour un sous-ensemble de la liste des 4717 spécificités, mais à partir de quatre nouvelles listes de spécificités. Ces listes de spécificités sont établies après comparaison de la liste de fréquence des lemmes d'un sous-corpus à la liste de fréquence des lemmes d'un sous-corpus extrait du journal *Le Monde* et qui respecte le rapport de 1 à 10 (Cf. annexe 16 pour les détails pratiques). Etant donné que la comparaison se fait avec un corpus de langue générale, les nouvelles listes de spécificités reflètent la thématique du domaine. Même si les sous-corpus ont tous leur propre liste de spécificités particulières, les quatre listes de spécificités par sous-corpus ont certainement des spécificités en commun tout comme elles comprennent des spécificités de la liste de base.

Comme notre but ici est de vérifier les conclusions des analyses de base, cette partie consacrée aux analyses détaillées par sous-corpus s'intéressera principalement à celui des normes. En effet, celle-ci sont censées être prescriptives et normatives. La question se pose donc de savoir si la corrélation négative entre le rang de monosémie et le rang de spécificité se maintient dans le sous-corpus des normes ou non. En d'autres mots, est-ce que la corrélation deviendra positive et est-ce que les mots les plus spécifiques dans les normes seront les plus homogènes sémantiquement ? Certes, les observations faites à partir des résultats des analyses

de régression (8.2.1) apportent une certaine réponse à la question, mais elles demandent à être complétées et mises au point (8.2.2).

8.2.1 Observations

Pour les observations concernant les analyses de régression détaillées par sous-corpus, nous reprenons le fil conducteur des analyses par classe lexicale, à savoir les coefficients de corrélation (8.2.1.1) et les résultats des analyses de régression : le pourcentage de R² et les variables significatives (8.2.1.2).

8.2.1.1 Les coefficients de corrélation

Les coefficients de corrélation Pearson des quatre sous-corpus (Cf. tableau 8.6) sont tous négatifs[216], confirmant ainsi de nouveau la corrélation négative entre le rang de monosémie (technique) et le rang de spécificité. De même, la corrélation pour le rang de monosémie technique est généralement plus faible que pour le rang de monosémie.

	coefficient de corrélation Pearson : rang de monosémie (technique) ~ rang de spécificité
mots4717	rangs 1-4717
mono	-0,71
mono tech	-0,65
revues 3025	rangs 1-3025
mono	-0,65
mono tech	-0,51
fiches 2650	rangs 1-2650
mono	-0,67
mono tech	-0,57
normes 1757	rangs 1-1757
mono	-0,69
mono tech	-0,62
manuels 1825	rangs 1-1825
mono	-0,72
mono tech	-0,62

Tableau 8.6 Corrélations par sous-corpus

[216] Ils sont tous statistiquement significatifs.

Les normes et les manuels se caractérisent par les meilleures corrélations, tant pour le rang de monosémie que pour le rang de monosémie technique. Ils sont comparables aux corrélations des 4717 spécificités de base. Les revues en revanche ont les coefficients de corrélation les plus faibles. D'ailleurs, pour les revues comme pour les manuels, on observe une chute importante si on passe du coefficient du rang de monosémie à celui du rang de monosémie technique.

Comme le sous-corpus des fiches se constitue de fiches techniques souvent courtes, il est plus sensible que les autres sous-corpus, aux frontières de documents (Cf. chapitre 5). Dès lors, afin de limiter l'effet de la transgression des frontières de documents, nous procédons aussi à l'analyse des cooccurrences dans une fenêtre plus limitée de deux mots à gauche et à droite. Toutefois, les résultats des analyses de régression pour les 1503 spécificités (de fich2) sont plutôt décevants. Les faibles coefficients de corrélation de -0,42 (rang de monosémie) et de -0,27 (rang de monosémie technique) sont liés aux rangs et aux degrés de monosémie particuliers de ces 1503 spécificités et sont dus au fait que la désambiguïsation dans une fenêtre d'observation très restreinte est nécessairement moins bonne.

8.2.1.2 Les résultats des analyses de régression : R^2 et variables significatives

Les résultats des analyses de régression simple et multiple (Cf. tableau 8.7) confirment les tendances des coefficients de corrélation et apportent des précisions complémentaires.

Comme les analyses par sous-corpus sont conduites sur quatre nouvelles listes de spécificités, il est impossible de situer visuellement les résultats sur la visualisation de base des 4717 spécificités. Pour les analyses de régression multiple par sous-corpus, la vérification des VIF des variables indépendantes requiert la suppression du degré de spécificité (log_LLR) et le remplacement du rang de fréquence générale par l'écart des rangs de fréquence (Cf. analyse de base). Par conséquent, la variable combinée (log_LLR et écart des rangs de fréquence) ne sera pas intégrée dans les modèles de régression multiple.

	simple R^2		**multiple R^2**	VI → rvfq2 remplacé par écart ; partout log_LLR supprimé (VIF)
	partout hétéroscéd.	diff. de R^2		
mots4717				
mono	51,57%	8,83%	80,65%	rvfq1 ; rvspec ; long ; *nbr_claslex*
mono tech	42,74%		75,31%	rvfq1 ; long ; *fqabs1 ; rvspec* ; *nbr_claslex*
revues 3025				
mono	42,28%	15,83%	70,75%	rvfq1; long ; *fqabs1* ; rvspec ; *écart*
mono tech	26,45%	chute !!	58,01%	rvfq1; long ; *fqabs1* ; *rvspec*
fiches 2650				
mono	45,20%	12,32%	69,03%	rvfq1; rvspec ; long ; *fqabs1* ; *fqabs2*
mono tech	32,88%		59,40%	rvfq1; *fqabs1* ; *rvspec* ; long
normes 1757				
mono	47,60%	8,06%	82,67%	rvfq1; rvspec ; long
mono tech	39,54%		78,37%	rvfq1; écart; long ; *nbr_claslex*
manuels 1825				
mono	53,14%	13,98%	73,47%	rvfq1; long ; *écart*; *fqabs1*
mono tech	39,16%	chute !!	61,46%	rvfq1; *fqabs1* ; long ; *écart* ; *rvspec*

Tableau 8.7 Résultats des analyses de régression par sous-corpus

Les tests de Goldfeld-Quandt mettent en évidence qu'il y a de l'hétéroscédasticité dans les quatre sous-corpus, que les visualisations des spécificités confirment d'ailleurs (Cf. annexe 16 : figures A16.1 à A16.8). Dans les analyses de régression simple, les normes et les manuels se caractérisent par les pourcentages de variation expliquée R^2 les plus élevés. Dans ces deux sous-corpus, plutôt prescriptifs, la variation du rang de spécificité permet de rendre compte de la variation du rang de monosémie et de celle du rang de monosémie technique. Dans les revues, en revanche, les pourcentages de R^2 sont les plus faibles[217], tant dans les analyses de régression simple que de régression multiple. Les meilleurs pourcentages de R^2 dans les analyses multiples s'observent dans les normes, où la différence entre le pourcentage pour le rang de monosémie et pour le rang de monosémie technique n'est que de 4%. Par contre, les manuels et les revues présentent une différence très importante de 12% dans les analyses multiples et de 14 à 15% dans les analyses

[217] Les pourcentages de R^2 de fich2 (fenêtre de 2 mots à gauche et 2 à droite dans les fiches) sont encore plus faibles : ils s'élèvent à 17% et 7% dans les analyses de régression simple et à 30% et 17% dans les analyses de régression multiple.

simples. Il est à remarquer que le sous-corpus des manuels, qui a un bon pourcentage de R^2 dans les analyses simples, révèle un pourcentage de R^2 plutôt moyen dans les analyses multiples.

En ce qui concerne les variables indépendantes significatives, le rang de fréquence technique reste la variable la plus significative, dans la mesure où elle explique le plus de variation du rang de monosémie et du rang de monosémie technique. Les spécificités les plus fréquentes d'un sous-corpus sont donc les moins homogènes sémantiquement. La longueur se maintient également partout : les spécificités les plus courtes sont toujours les moins homogènes sémantiquement, quel que soit le sous-corpus.

Le rang de spécificité est significatif dans les revues et les fiches : il affiche une corrélation négative avec le rang de monosémie et une corrélation positive avec le rang de monosémie technique, à l'instar des résultats pour les 4717 spécificités de base (Cf. chapitre 7). Le rang de spécificité se caractérise aussi par une corrélation négative avec le rang de monosémie dans les normes et par une corrélation positive avec le rang de monosémie technique dans les manuels. Si le rang de spécificité ne figure pas parmi les variables significatives, l'écart des rangs de fréquence, qui indique la plus ou moins grande technicité des spécificités, y a sa place. Tel est le cas du rang de monosémie dans les manuels et du rang de monosémie technique dans les normes.

8.2.2. Interprétations et mises au point

8.2.2.1 Interprétations linguistiques et explications quantitatives

Les résultats des analyses de régression par sous-corpus confirment donc la corrélation négative entre le rang de spécificité et le rang de monosémie (et respectivement le rang de monosémie technique). Dans les normes et les manuels, les deux sous-corpus les plus prescriptifs et les plus normatifs, cette corrélation négative est la plus forte. Les résultats sont d'autant plus concluants pour réfuter la thèse des monosémistes, que le degré de technicité des normes et des manuels est plus élevé. En effet, si on adopte le point de vue traditionnel des monosémistes, on ne s'attendrait pas à une telle corrélation négative dans le sous-corpus des normes. En d'autres mots, on ne s'attendrait pas du tout à ce que les spécificités les plus spécifiques, ou les plus typiques du domaine de spécialité, soient les plus hétérogènes sémantiquement, bien au contraire.

En plus, les normes se caractérisent par une différence limitée entre les pourcentages de variation expliquée R^2 pour le rang de monosémie et pour le rang de monosémie technique, à savoir une différence de 4% dans les analyses de régression multiple et de 8% dans les analyses de régression simple. Tout compte fait, de par leur nature,

les normes sont des textes hautement techniques. Par contre, les revues et les manuels sont les plus sensibles à la mesure de monosémie technique, parce que ces deux sous-corpus affichent les différences les plus importantes, 12% dans les analyses de régression multiple et presque 16% et 14% dans les analyses de régression simple.

L'explication linguistique de cette sensibilité réside principalement dans la présence de cc généraux, qui sont responsables de la chute des degrés de monosémie technique et donc des modifications dans la répartition des rangs de monosémie technique. Ils entraînent à leur tour des modifications dans les coefficients de corrélation et dans les pourcentages de variation expliquée pour le rang de monosémie technique. La comparaison quantitative montre que les revues et les manuels ont effectivement plus de cc plus fréquents et plus de cc plus généraux : la moyenne de la fréquence moyenne pondérée est effectivement plus faible dans les revues (0,82) et dans les manuels (0,90) que dans les normes (0,95). Cela s'explique par le fait que les revues et les manuels ont un niveau de vulgarisation plus élevé, que ce sont des sources plus accessibles dans la mesure où les revues sont plus descriptives et les manuels constituent des documents à visée didactique (Cf. chapitre 3). Les différents niveaux de normalisation et de vulgarisation des quatre sous-corpus (Cf. tableau 8.8) sont donc à la base des résultats des analyses de régression détaillées par sous-corpus.

	vulgarisation +	vulgarisation -
normalisation +	manuels	normes
normalisation -	revues	fiches

Tableau 8.8 Niveaux de normalisation et de vulgarisation des sous-corpus

L'explication linguistique des pourcentages de R^2 élevés et faibles en fonction des niveaux de normalisation et de vulgarisation des quatre sous-corpus, est confirmée par l'explication quantitative à partir de la fréquence absolue dans le corpus général. Les spécificités dans les fiches et dans les normes, c'est-à-dire dans les sous-corpus les moins vulgarisateurs, sont en moyenne peu fréquentes dans le corpus général : la moyenne de leur fréquence absolue dans le corpus général est de 59 et de 97 respectivement. Par contre, les spécificités dans les revues sont plus fréquentes dans le corpus général (moyenne de 201), ce qui se traduit par des pourcentages de R^2 plus faibles. Si les spécificités d'un sous-corpus sont, en moyenne, plus fréquentes dans le corpus de référence de langue générale (Cf. revues), l'explication de la variation du rang de monosémie sera moins bonne pour les spécificités les moins spécifiques, d'où les pourcentages plus faibles de R^2. Ainsi, on constate que les

spécificités des normes sont, en moyenne, moins fréquentes dans le corpus de référence de langue générale et qu'elles se prêtent mieux à l'explication de la variation du rang de monosémie par le rang de spécificité, donnant lieu à des pourcentages plus élevés de R^2.

Les observations et les interprétations pour les spécificités des quatre sous-corpus confirment donc les conclusions des analyses de base pour les 4717 spécificités. En plus, elles réfutent l'hypothèse traditionnelle selon laquelle les spécificités des normes seraient homogènes sémantiquement.

8.2.2.2 Les normes par rapport aux trois autres sous-corpus

Etant donné que les normes constituent un sous-corpus très intéressant du point de vue de la thèse des monosémistes, nous aimerions approfondir l'analyses de ce sous-corpus clé.

Dans la section précédente (Cf. 8.2.2.1), les spécificités des normes ont été déterminées par la comparaison du sous-corpus technique des normes à un corpus de langue générale du journal *Le Monde*. Ces 1757 spécificités plutôt thématiques reflètent donc le domaine de spécialité. Dans cette section, une nouvelle liste de spécificités des normes sera établie à partir de la comparaison du sous-corpus technique des normes aux trois autres sous-corpus techniques, qui sont intégrés dans un nouveau corpus de référence « rfm » (revues, fiches, manuels). Cette nouvelle liste de spécificités des normes (1471 spécificités) contiendra moins de spécificités typiques du domaine, telles que *usinage*, étant donné que ces spécificités apparaissent aussi et peut-être même plus souvent dans les trois autres sous-corpus, qui constituent le nouveau corpus de référence. La nouvelle liste de spécificités comprendra surtout des spécificités propres aux particularités stylistiques des normes par rapport aux autres sous-corpus, par exemple *sécurité, autorité*, bien que les spécificités thématiques plus fréquentes dans les normes y figurent aussi. Les principes de génération de cette nouvelle liste de spécificités sont expliqués en annexe (Cf. annexe 16).

Compte tenu de la thèse des monosémistes, on pourrait avancer ici aussi l'hypothèse classique que les mots qui sont spécifiques dans les normes (1471), par rapport aux trois autres sous-corpus techniques, sont plus monosémiques ou plus homogènes sémantiquement. Cette nouvelle hypothèse fait écho à l'hypothèse (Cf. 8.2.2.1) que les mots spécifiques apparaissant dans les normes, si on compare les fréquences dans les normes à celles d'un corpus de langue générale, sont plus monosémiques ou plus homogènes sémantiquement. Rappelons que les résultats des analyses statistiques de régression ont infirmé cette hypothèse. Afin de vérifier la nouvelle hypothèse concernant les spécificités des normes (1471), nous implémenterons les variables dépendantes pour cette nouvelle liste, c'est-à-dire le rang de monosémie et

le rang de monosémie technique. Nous compléterons aussi la liste des variables indépendantes[218] en introduisant le rang de spécificité, le degré de spécificité (log_LLR), le rang de fréquence dans les normes (rang_v_freq1), la fréquence absolue dans les normes (fqabs1), le rang de fréquence dans le corpus de référence rfm (rang_v_freq2), la fréquence absolue dans le corpus de référence rfm (fqabs2) et la longueur des spécificités.

Le tableau ci-dessus (Cf. tableau 8.9) visualise la corrélation négative entre le rang de spécificité et le rang de monosémie de ces 1471 nouvelles spécificités et infirme donc l'hypothèse de la monosémie des spécificités dans les normes. En effet, les mots les plus spécifiques dans les normes, déterminés par la comparaison des normes aux autres trois sous-corpus, sont les plus polysémiques. Toutefois, la tendance négative est moins claire que celle des spécificités des normes par rapport au corpus de langue générale, parce que le coefficient de corrélation est moins important (-0,64 versus -0,69) et que les points sont plutôt dispersés (Cf. annexe 16 : figure A16.9).

```
        Pearson's product-moment correlation

data:  rang_v_mono_0.9999 and rang_v_spec
t = -32.4926, df = 1469, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6754426 -0.6159023
sample estimates:
       cor
-0.6466564
```

Tableau 8.9 Corrélation : rang de monosémie ~ rang de spécificité : norm_rfm

Le tableau suivant (Cf. tableau 8.10) visualise les résultats des analyses de régression simple et multiple pour la nouvelle liste de spécificités des normes (norm_rfm), établie par la comparaison des normes aux trois autres sous-corpus techniques (revues, fiches, manuels). Ce tableau 8.10 rappelle également, à titre d'information, les résultats pour la liste précédente de spécificités des normes (norm_lm), spécifiques dans le sous-corpus des normes par rapport au sous-corpus de référence de langue générale.

[218] Les variables de la classe lexicale et du nombre de classes lexicales ne sont pas implémentées.

	simple R^2		multiple R^2	variables indépendantes
	partout hétéroscéd.	diff. de R^2		
norm_lm 1751				
mono	47,60%	8,06%	82,67%	rvfq1; rvspec ; long
mono tech	39,54%		78,37%	rvfq1; écart; long ; *nbr_claslex*
norm_rfm 1471				
mono	41,78%	11,44%	80,76%	rvfq1
mono tech	30,26%		69,35%	rvfq1; *rvspec*

Tableau 8.10 Résultats des analyses de régression : norm_lm et norm_rfm

D'abord, la comparaison quantitative indique que, globalement, les pourcentages de R^2 sont plus faibles dans norm_rfm que dans norm_lm, tant pour les analyses simples que pour les analyses multiples. En ce qui concerne les variables indépendantes, il est à remarquer que, pour norm_rfm, la vérification des VIF exclut le log_LLR et le rang de fréquence dans le corpus de référence rfm (rvfq2). Seul le rang de fréquence dans les normes est significatif, complété par le rang de spécificité pour le rang de monosémie technique. Dans les analyses de base (Cf. chapitre 7), un pourcentage plutôt faible de R^2 s'explique par la présence de plus de mots généraux (fréquents dans le corpus de référence de langue générale). La comparaison de la moyenne de la fréquence absolue dans le corpus général (dans le sous-corpus du journal *Le Monde*, qui a servi de corpus de référence pour norm_lm) pour les 1751 mots de norm_lm et pour les 1471 mots de norm_rfm permet effectivement de constater que la moyenne de fréquence absolue dans le corpus général dans norm_rfm est plus élevée (158) que celle dans norm_lm (97). Il est clair que la présence de plus de mots plus généraux, c'est-à-dire plus fréquents dans le corpus général, diminue de nouveau le pouvoir explicatif des modèles de régression simple et multiple, en l'occurrence pour la liste des 1471 spécificités des normes, comparées aux autres sous-corpus.

Par ailleurs, la comparaison qualitative de ces deux listes de spécificités, norm_lm et norm_rfm, permet d'identifier un groupe commun de 1130 spécificités, spécifiques dans les deux listes. Il s'ensuit que 627 spécificités sont typiques des normes par rapport au corpus général. Ce sont donc des spécificités proprement thématiques, qui ne figurent pas dans norm_rfm (Cf. tableau 8.11). Il reste 341 spécificités typiques des normes par rapport aux trois autres sous-corpus. Comme ces 341 ne figurent pas dans la liste norm_lm, elles reflètent des particularités stylistiques du sous-corpus des normes, par rapport aux trois autres sous-corpus techniques relevant du même domaine thématique (Cf. tableau 8.12). D'ailleurs, les 341 spécificités stylistiques caractéristiques de norm_rfm et absentes de norm_lm, sont plus générales (moyenne

de fréquence absolue dans le corpus général de 312) que les 627 spécificités thématiques, typiques de norm_lm (moyenne de 70).

outil	*équiper*	*copeau*
type	*mécanique*	*hydraulique*
usinage	*automatique*	*permettre*
pièce	*meule*	*vibration*
vitesse	*mm*	*surface*
matériau	*montage*	*machine-outil*
rotation	*fraise*	*fabrication*
système	*m*	*liquide*
broche	*diamètre*	*axe*
serrage	*contrôle*	*possible*

Tableau 8.11 Spécificités thématiques les plus spécifiques dans norm_lm (627)

pas	*ex*	*droit*
national	*publier*	*journal*
texte	*mandat*	*contrat*
f	*travailleur*	*territoire*
proposition	*titre*	*amendement*
social	*parlement*	*statut*
autorité	*feu*	*acte*
public	*déclarer*	*employeur*
pays	*décision*	*union*
conseil	*avis*	*fonctionnaire*

Tableau 8.12 Spécificités stylistiques les plus spécifiques dans norm_rfm (341)

8.2.2.3 Importance de la spécificité par source

Finalement, nous nous interrogeons sur l'importance de la spécificité selon la source, c'est-à-dire du rang de spécificité par sous-corpus, dans le but de vérifier son impact sur l'homogénéité sémantique dans le corpus entier.

On pourrait effectivement avancer l'hypothèse (d'inspiration classique) que les mots qui sont spécifiques dans les normes, sont en même temps plus monosémiques ou plus homogènes sémantiquement dans le corpus technique entier, dans la mesure où ils sont « imposés »[219]. Le but de cette section est donc de vérifier si les spécificités thématiques des normes, spécifiques dans les normes par rapport à un corpus de

[219] Cela n'empêche pas que ces mots s'utilisent parfois dans un sens autre que celui prévu au départ par ISO.

langue générale, sont (plus) monosémiques, lorsqu'elles sont employées dans le corpus technique entier relevant du même domaine thématique. On peut se demander également si les spécificités des autres sous-corpus sont susceptibles d'influencer le rang de monosémie (technique) dans le corpus entier. En général, la question se pose donc de savoir si la spécificité de la source a un impact statistiquement significatif sur le rang de monosémie et sur le rang de monosémie technique dans le corpus entier.

Les analyses de régression multiple visant à répondre à cette question, prennent en considération les spécificités de l'intersection des 5 listes, donc les spécificités qui sont communes aux 4 sous-corpus et au corpus entier. Ces 440 spécificités sont généralement plutôt spécifiques dans le corpus entier (rangs de spécificité dans le corpus entier inférieurs à 2275). Le tableau 8.13 ci-dessous visualise les variables dépendantes et indépendantes pour les 5 mots les plus spécifiques du corpus technique entier, à savoir le rang de spécificité, le rang de monosémie et le rang de monosémie technique, et en plus, le rang de spécificité dans les revues (revu_v_spec), les fiches (fich_v_spec), les normes (norm_v_spec) et les manuels (manu_v_spec). Les quatre dernières colonnes indiquent donc les rangs de spécificité par sous-corpus ou la spécificité par source.

rang_ v_spec	spécificité	rang_ v_mono	rang_ v_mono_tech	revu_ v_spec	fich_ v_spec	norm_ v_spec	manu_ v_spec
1	*machine*	4717	4712	1	2	1	18
2	*outil*	4715	4711	4	3	25	2
3	*usinage*	4614	4513	2	6	39	3
4	*pièce*	4670	4607	3	4	47	9
5	*mm*	4079	3806	5	1	142	23

Tableau 8.13 Rangs de spécificité par sous-corpus

- *Coefficients de corrélation*

Si l'hypothèse traditionnelle formulée ci-dessus se vérifie, le rang de spécificité dans les normes (norm_v_spec) devrait se caractériser par une corrélation positive avec le rang de monosémie (et avec le rang de monosémie technique), les mots les plus spécifiques dans les normes (rangs de spécificité près de 1) étant les plus monosémiques dans le corpus entier (rangs de monosémie près de 1). Or, le tableau ci-dessous (Cf. tableau 8.14) montre une corrélation négative entre, d'une part, le rang de spécificité dans les différents sous-corpus, y compris le sous-corpus des normes, et d'autre part, le rang de monosémie (technique) dans le corpus entier. Cette corrélation négative est plus faible dans les manuels et les normes, ce qui reflète quand même leur caractère prescriptif. Pourtant, selon la thèse traditionnelle des monosémistes, celui-ci aurait dû se manifester à travers une corrélation positive.

	rang_v_spec	rang_v_mono_0.9999	rang_v_mono_WLLR_0.9999
rang_v_spec	1.0000000	-0.7219056	-0.6720318
rang_v_mono_0.9999	-0.7219056	1.0000000	0.9826105
rang_v_mono_WLLR_0.9999	-0.6720318	0.9826105	1.0000000
revu_v_spec	0.8362675	-0.5529526	-0.5073617
fich_v_spec	0.7697371	-0.5498127	-0.4980305
norm_v_spec	**0.4085340**	**-0.4186508**	**-0.4231335**
manu_v_spec	**0.6104033**	**-0.4007913**	**-0.3729555**

Tableau 8.14 Corrélations des rangs de spécificité par sous-corpus

Le tableau 8.14 montre également de faibles coefficients de corrélation entre, d'une part, le rang de spécificité dans le corpus entier (rang_v_spec) et, de l'autre, le rang de spécificité dans les manuels (0,61) et dans les normes (0,40), contrairement aux meilleures corrélations dans les revues (0,83) et les fiches (0,76). Cette différence s'explique par la part plus importante qu'occupent les revues dans le corpus entier (45%) par rapport aux trois autres sous-corpus (fiches 17%, normes 16% et manuels 22%). En plus, les listes de spécificités et plus particulièrement les valeurs du rang de spécificité dans les revues et les fiches ressemblent plus à la liste de spécificités et aux valeurs du rang de spécificité dans le corpus entier, en raison des particularités thématiques et stylistiques des manuels et des normes (Cf. 8.2.2.2).

- *Résultats des analyses de régression multiple : variables significatives*

Les modèles de régression multiple du tableau 8.15 font intervenir, comme variable dépendante, les rangs de monosémie et de monosémie technique dans le corpus entier, et comme variables indépendantes, (1) le rang de spécificité dans le corpus entier et les 4 rangs de spécificité dans les sous-corpus et (2) les 4 rangs de spécificité dans les sous-corpus exclusivement. Ces modèles de régression multiple permettent donc d'identifier la variable indépendante la plus importante pour expliquer la variation du rang de monosémie (technique) dans le corpus entier. La vérification des VIF des 5 et, respectivement, des 4 variables indépendantes ne soulève aucun problème de multicolinéarité.

	R² et variables indépendantes significatives	
Variable dépendante	(1) avec rang_v_spec	(2) sans rang_v_spec
Rang de monosémie	54,42%	45,08%
	rang_v_spec **norm_v_spec** *revu_v_spec* *manu_v_spec*	4 VI **norm_v_spec** (valeur p : 2,79e-13)
Rang de monosémie technique	48,18%	39,82%
	rang_v_spec **norm_v_spec** *revu_v_spec*	4 VI **norm_v_spec** (valeur p : 9,92e-14)

Tableau 8.15 Régression multiple : rangs de spécificité par sous-corpus

Le tableau 8.15 montre que, dans le modèle avec les 5 variables indépendantes (1), le rang de spécificité dans le corpus entier est la variable la plus significative (corrélation négative). La deuxième variable la plus importante est le rang de spécificité dans les normes (corrélation négative). En dépit de la moins bonne corrélation individuelle, cette variable (norm_v_spec) est tout de même significative dans le modèle de régression multiple à 5 variables indépendantes. Dans le modèle de régression multiple faisant intervenir uniquement les 4 rangs de spécificité dans les sous-corpus, les 4 variables indépendantes sont toutes significatives (corrélations négatives). Le rang de spécificité dans les normes étant la variable la plus significative malgré sa faible corrélation individuelle avec le rang de monosémie (technique) (Cf. tableaux 8.14 et 8.15). Il est à remarquer que le rang de spécificité dans les revues est faiblement significatif dans le modèle à 5 variables indépendantes. En effet, sa corrélation avec la variable dépendante est positive dans le modèle à 5 variables indépendantes en dépit de sa corrélation négative individuelle avec le rang de monosémie et avec le rang de monosémie technique, qui est plutôt bonne.

- *Explications et interprétations*

Comment expliquer ou interpréter la significativité importante du rang de spécificité dans les normes, ainsi que la corrélation positive (faiblement significative) du rang de spécificité dans les revues ? En fait, ces contradictions apparentes n'en sont pas, mais elles reposent sur des effets de complémentarité des variables indépendantes.

Dans le modèle de régression multiple faisant intervenir les 5 rangs de spécificité, le rang de spécificité dans le corpus entier explique clairement la majeure partie de la variation du rang de monosémie (technique). La variation restante est expliquée par les autres variables significatives, en l'occurrence le rang de spécificité dans les normes (corrélation négative) et dans les revues (corrélation positive). Ces deux variables sont donc complémentaires par rapport au rang de spécificité dans le corpus entier. Comme les valeurs du rang de spécificité dans les normes se distinguent le plus clairement des valeurs du rang de spécificité dans le corpus entier, le rang de spécificité dans ce sous-corpus est le plus complémentaire par rapport au rang de spécificité dans le corpus entier. Par conséquent, il est le plus significatif dans le modèle de régression multiple faisant intervenir les 5 variables indépendantes.

En effet, certaines spécificités se comportent différemment dans les normes que dans le corpus entier, ayant un rang de spécificité particulier (Cf. figure 8.1). Dans la partie inférieure gauche du nuage de points, les spécificités sont spécifiques dans le corpus entier, mais moins spécifiques dans les normes (points en bleu et en mauve). Ces spécificités sont plus monosémiques dans le corpus entier qu'on ne penserait

compte tenu de leur rang de spécificité assez spécifique dans le corpus entier (selon la corrélation négative). Néanmoins, leur spécificité plus limitée dans les normes permet de mieux rendre compte de leur rang de monosémie dans le corpus entier. Il en va de même pour les spécificités en haut au milieu (Cf. figure 8.1) : elles sont moyennement spécifiques dans le corpus entier, mais plus spécifiques dans les normes (en jaune), ce qui explique leur hétérogénéité sémantique dans le corpus entier (rangs de monosémie près de 4700), d'après la corrélation négative.

Dans le modèle multiple avec les 4 rangs de spécificité dans les sous-corpus, la significativité plus importante du rang de spécificité dans les normes s'explique par l'interaction avec les autres variables indépendantes qui peuvent altérer la pertinence de son pouvoir explicatif.

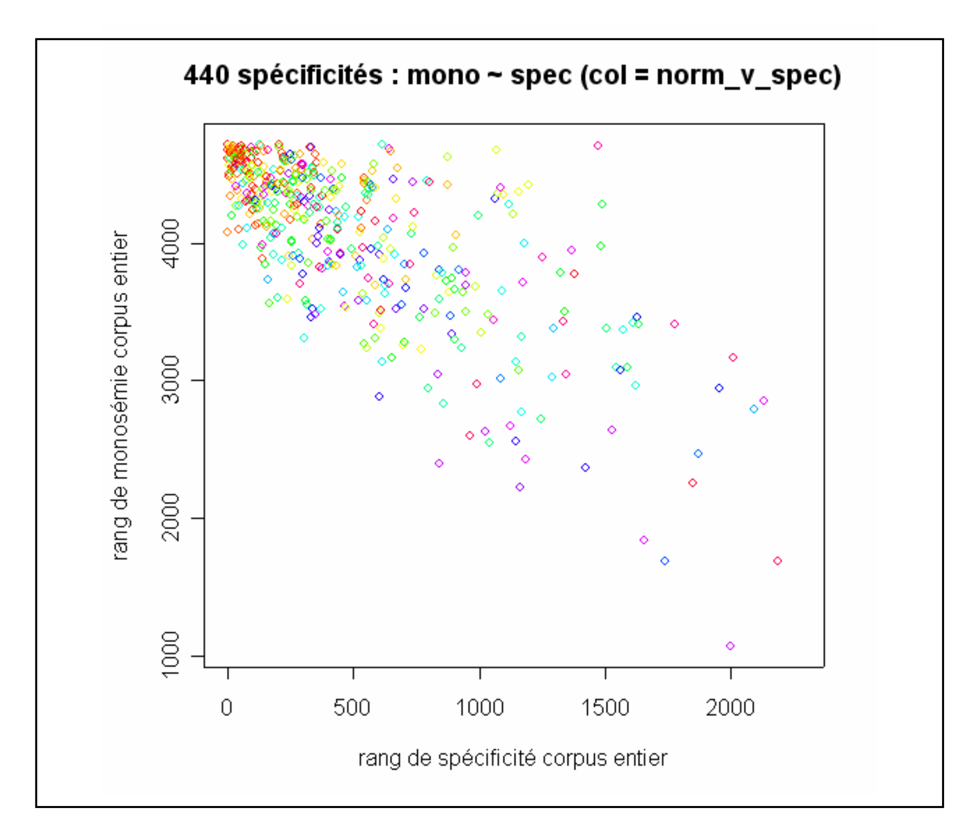


Figure 8.1 Régression simple : rang de spécificité (dans les normes) en couleur

Le rang de spécificité dans les revues est également complémentaire par rapport au rang de spécificité dans le corpus entier. La corrélation positive pourrait donc s'interpréter en fonction du pouvoir explicatif complémentaire du rang de spécificité dans les revues, en particulier pour les mots à résidus importants. Ceux-ci illustrent

moins bien le pouvoir explicatif du rang de spécificité dans le corpus entier. La figure ci-dessous (Cf. figure 8.2) montre effectivement que les couleurs du rang de spécificité dans les revues suivent bien les rangs de spécificité dans le corpus entier, à quelques exceptions près. Ces exceptions font l'objet du pouvoir explicatif complémentaire du rang de spécificité dans les revues.

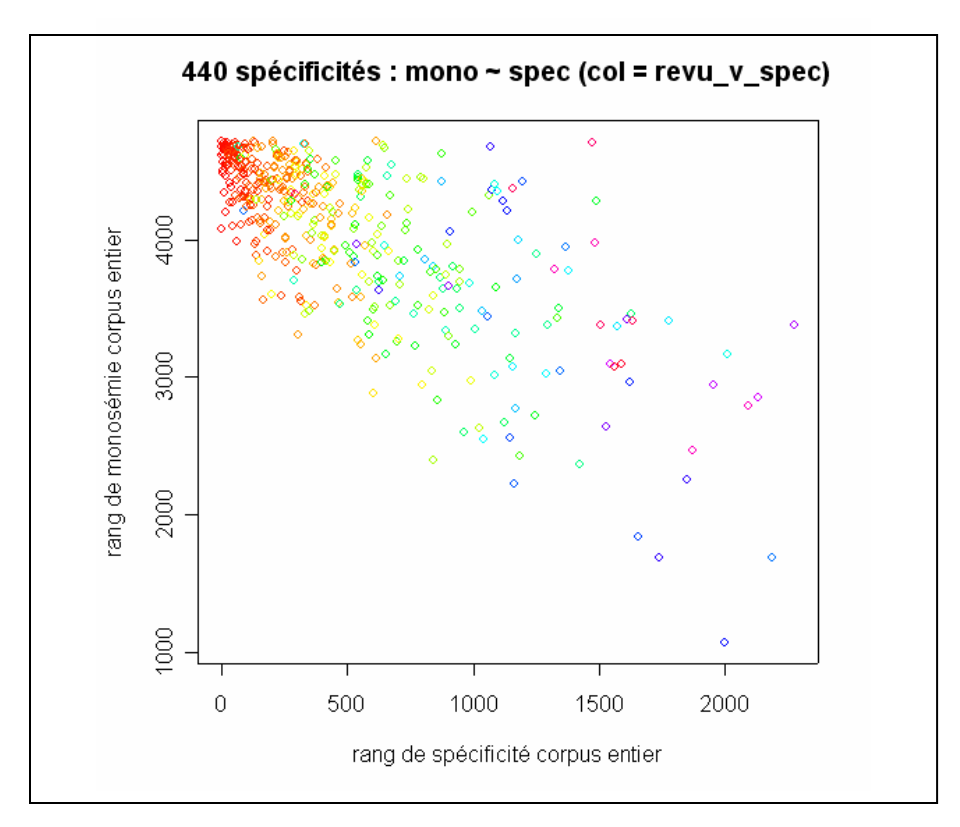


Figure 8.2 Régression simple : rang de spécificité (dans les revues) en couleur

En conclusion, les analyses de corrélation et de régression multiple de cette section (8.2.2.3) nous ont permis d'infirmer l'hypothèse selon laquelle les mots qui sont spécifiques dans les normes (par rapport à un corpus de référence de langue générale), seraient plus monosémiques dans le corpus entier. En effet, la corrélation négative entre le rang de spécificité dans les normes et le rang de monosémie dans le corpus entier révèle le contraire. En plus, dans un modèle de régression multiple, qui intègre tous les rangs de spécificité (dans le corpus entier et dans les sous-corpus), le rang de spécificité dans les normes s'avère particulièrement significatif pour expliquer la variation du rang de monosémie dans le corpus entier et se caractérise aussi par une corrélation négative.

8.3 CONCLUSION DES ANALYSES DÉTAILLÉES

Les analyses de régression détaillées nous ont permis de vérifier la corrélation entre, d'une part, le rang de spécificité et, de l'autre, le rang de monosémie et le rang de monosémie technique, par classe lexicale (substantifs, adjectifs, verbes et adverbes) et par sous-corpus (revues, fiches, normes, manuels).

Pour les différentes classes lexicales de la liste des 4717 spécificités, nous avons réfuté l'hypothèse d'inspiration classique que les mots les plus spécifiques d'une classe lexicale sont les plus monosémiques. Les analyses de régression simple et multiple pour les quatre classes lexicales montrent effectivement le contraire, c'est-à-dire une corrélation négative entre le rang de spécificité et le rang de monosémie (technique). Au moment dit, les mots les plus spécifiques d'une classe lexicale sont les plus polysémiques à l'intérieur de cette classe lexicale. Il s'avère aussi que la classe lexicale des substantifs est celle qui représente le mieux cette corrélation négative. Dès lors, c'est la classe par excellence qui corrobore le pouvoir explicatif du rang de spécificité, ce qui remet définitivement en question la thèse monosémiste. La classe lexicale des adverbes au contraire, illustre moins bien le pouvoir explicatif du rang de spécificité.

Nous avons vu que l'explication quantitative des analyses de base, qui apporte une solution au problème de l'hétéroscédasticité (Cf. chapitre 7), s'applique également aux résultats des analyses détaillées par classe lexicale. Si un sous-ensemble de spécificités comprend plus de spécificités générales (ou fréquentes dans le corpus de langue générale), il se prête moins bien à une analyse suivant les modèles de régression simple et multiple. Pour les spécificités les plus générales, le modèle de régression n'est guère satisfaisant, parce qu'il donne lieu à l'hétéroscédasticité et/ou à des pourcentages de variation expliquée R^2 plutôt faibles. Une explication linguistique en termes de propriétés syntaxiques et collocationnelles permet en revanche de corroborer les résultats et les conclusions des analyses détaillées par classe lexicale. Rappelons que les substantifs se caractérisent par des mécanismes collocationnels particuliers et très puissants, parce qu'ils sont désambiguïsés par les adjectifs qui les modifient ou les qualifient. Pour les adverbes, au contraire, les mécanismes collocationnels et les critères de sélection sont moins clairs et moins restrictifs. Le pouvoir désambiguïsateur de leurs cooccurrents, moins nombreux et moins forts, a par conséquent un impact considérable sur les rangs de monosémie et dès lors, sur les résultats des analyses de régression.

Pour les analyses de régression détaillées par sous-corpus, nous avons formulé trois hypothèses. Elles concernent surtout sur le sous-corpus des normes, étant donné qu'il constitue le sous-corpus le plus intéressant et le plus concluant, toujours dans la perspective d'une réfutation de la thèse des monosémistes. Les questions que nous

avons posées sont, premièrement, les mots les plus spécifiques dans le sous-corpus des normes, si on compare les fréquences dans les normes à celles d'un corpus de langue générale, sont-ils les plus homogènes sémantiquement ? Deuxièmement, les mots les plus spécifiques dans les normes, si on compare les fréquences dans les normes à celles des trois autres sous-corpus techniques, sont-ils en même temps les plus homogènes sémantiquement ? Et, finalement, les mots les plus spécifiques dans les normes, si on compare les fréquences dans les normes à celles d'un corpus de langue générale, sont-ils les plus homogènes sémantiquement dans le corpus technique entier, relevant du même domaine de spécialité ?

Les résultats des différentes analyses de régression sont convergents et conduisent à réfuter ces trois hypothèses. En effet, les mots les plus spécifiques dans les normes, tant par rapport à un corpus de langue générale que par rapport aux trois autres sous-corpus techniques, ne sont pas les mots les plus homogènes sémantiquement. Au contraire, ils sont les plus hétérogènes. De plus, le sous-corpus des normes se distingue des revues et des manuels par la différence limitée entre les pourcentages de variation expliquée R^2 pour le rang de monosémie et pour le rang de monosémie technique. Les normes sont donc le sous-corpus le moins sensible aux effets de la mesure de monosémie technique. Les revues et les manuels sont plus sensibles à ces effets en raison de leur niveau de vulgarisation plus élevé et du nombre plus important de cc généraux, parce qu'il s'agit de sources plus accessibles, descriptives et didactiques.

Finalement, les analyses de régression multiple qui font intervenir les différents rangs de spécificité dans les quatre sous-corpus et dans le corpus entier, ont permis d'infirmer l'hypothèse que les mots les plus spécifiques dans les normes sont les plus homogènes sémantiquement dans le corpus entier. Qui plus est, le rang de spécificité dans les normes est une variable significative pour expliquer la variation du rang de monosémie dans le corpus entier.

Chapitre 9
Conclusions et perspectives

L'objectif principal de cette étude était de procéder à une analyse sémantique quantitative du vocabulaire spécifique d'un domaine technique, en l'occurrence le domaine des machines-outils pour l'usinage des métaux. Plus particulièrement, le but était de vérifier si et dans quelle mesure les unités lexicales spécifiques de ce domaine sont monosémiques ou polysémiques. A cet effet, nous avons adopté une double approche, quantitative et scalaire. Le deuxième objectif, corollaire du premier, était de développer une mesure du degré de monosémie, dans le but de quantifier l'analyse sémantique et de procéder à des analyses de régression.

Le point de départ de notre étude était le constat que de nombreuses études récentes ont remis en question l'idéal de monosémie ou la thèse monosémiste de l'approche traditionnelle. Ces études ont clairement démontré qu'il existe de la polysémie dans un corpus de langue spécialisée, mais se sont limitées à l'analyse sémantique de quelques unités lexicales. Dans notre étude, nous nous sommes également engagée dans la voie de l'analyse descriptive. Toutefois, nous y avons rajouté la dimension de l'analyse quantitative et statistique, qui permet une analyse sémantique à plus grande échelle.

Dans un premier temps, nous avons dégagé dans l'état de la question les idées fondamentales qui permettent de reformuler la thèse monosémiste en une question mesurable. Ainsi, la question principale était celle de savoir si les mots les plus spécifiques du corpus technique sont effectivement les plus monosémiques. C'est cette question qui nous a fait recourir à une double approche quantitative, à savoir l'analyse des spécificités et l'analyse des cooccurrences. Finalement, la double approche quantitative a conduit à des analyses statistiques de régression, dont les résultats ont apporté des réponses tant aux questions de base et qu'aux questions détaillées, comme nous le rappelons brièvement ci-dessous.

Arrivée au terme de notre travail, nous mettrons en évidence les conclusions générales et les lignes de force de notre étude (9.1) et nous terminerons par des perspectives de recherches futures (9.2).

9.1 CONCLUSIONS GENERALES

- *La remise en question de la dichotomie traditionnelle*

L'état de la question nous a permis de situer le cadre théorique de notre étude, aussi bien au niveau de la langue spécialisée, qu'au niveau de l'analyse sémantique (*chapitre 1*). Ces deux niveaux se caractérisent par une dichotomie qui a été remise en question pour de nombreuses raisons, surtout par les partisans de l'approche descriptive.

En ce qui concerne la langue spécialisée, la dichotomie traditionnelle entre la langue générale et la langue spécialisée, ou entre le mot et le terme, ne s'est pas avérée appropriée pour l'analyse quantitative d'un corpus spécialisé. Selon la tradition, les mots font partie de la langue générale, tandis que les termes sont réservés aux langues spécialisées. Or, il est évident que le vocabulaire d'un corpus technique ne comprend pas uniquement des termes propres au domaine, mais également des mots du Vocabulaire Général d'Orientation Scientifique (VGOS). Ces derniers s'emploient dans plusieurs domaines scientifiques et techniques et leur sens est déterminé par les contextes spécialisés. En plus, le vocabulaire d'un corpus spécialisé comprend aussi des mots de la langue générale, tant des unités lexicales que des unités grammaticales. Par ailleurs, il s'est avéré que les termes voyagent non seulement d'un domaine à l'autre, mais également de la langue spécialisée à la langue générale et inversement. Ces diverses interactions et ces processus de nomadisation et de (dé)terminologisation, ainsi que l'absence d'un classement strictement binaire (« mot » versus « terme »), nous ont incitée à adopter la solution alternative d'une approche scalaire, autrement dit une approche par continuum, avec des unités lexicales considérées comme plus ou moins spécifiques du corpus technique.

Pour ce qui est de l'analyse sémantique, la langue spécialisée d'un corpus technique (ou d'un corpus relevant d'un domaine spécialisé), se caractérise idéalement, selon la vision traditionnelle, par la monosémie et la monoréférentialité. La polysémie serait évitée ou réduite à l'homonymie, qui fait intervenir deux ou plusieurs domaines spécialisés. Toutefois, des expérimentations récentes sur des corpus spécialisés ont démontré la présence indéniable de la polysémie dans la langue spécialisée, même à l'intérieur d'un seul domaine. Par ailleurs, les critères traditionnels permettant de distinguer la monosémie, la polysémie, l'homonymie et l'indétermination ne sont pas toujours fiables ni convergents.

- *La solution alternative : une approche scalaire à deux niveaux*

Il est clair que la dichotomie traditionnelle qui oppose la polysémie à la monosémie ne correspond pas à la dichotomie qui oppose la langue générale à la langue

spécialisée. Par conséquent, nous avons décidé d'adopter la solution alternative d'une approche scalaire, tant pour les unités lexicales spécifiques que pour leur analyse sémantique. Cette approche scalaire nous a amenée à situer les unités lexicales spécifiques du corpus technique sur un continuum de spécificité, ainsi que sur un continuum sémantique (de monosémie).

Etant donné la méthodologie pour laquelle nous avons opté, nous avons reformulé la thèse monosémiste en une question quantitative (*chapitre 2*). Si la thèse monosémiste devait se vérifier, elle aurait été particulièrement vraie pour les unités lexicales les plus spécifiques du corpus technique. Dès lors, la question s'est posée de savoir si les unités lexicales les plus spécifiques du corpus technique sont effectivement les plus monosémiques. La réponse à cette question revient à étudier la corrélation entre le continuum de spécificité, d'une part, et le continuum de monosémie, de l'autre. S'il y a une corrélation positive, donc si les unités les plus spécifiques sont effectivement les unités les plus monosémiques, la thèse monosémiste se vérifie. Sinon, elle est infirmée. Bien évidemment, pour implémenter les gradations tant de spécificité que de monosémie il a fallu une double analyse quantitative, au niveau des spécificités et au niveau sémantique.

- *Vers une double analyse quantitative : spécificités et cooccurrences*

Dans un premier temps, et après avoir expliqué la constitution du corpus technique et du corpus de référence de langue générale (*chapitre 3*), nous avons présenté et discuté les deux approches méthodologiques envisageables pour identifier les unités lexicales spécifiques du corpus technique (*chapitre 4*). Nous avons vu que les deux approches, c'est-à-dire le calcul des spécificités et la méthode des mots-clés, attribuent un degré de spécificité aux unités spécifiques identifiées. Compte tenu de la granularité des résultats et de l'efficacité technique, la méthode des mots-clés a été retenue comme la méthode la plus appropriée pour notre première analyse quantitative, l'analyse des spécificités. En opposant le corpus technique à un corpus de référence de langue générale, la méthode des mots-clés a permis de générer une liste de 4717 spécificités statistiquement significatives, spécifiques du corpus technique. L'indication du degré de spécificité a permis, par la suite, de les situer sur un continuum de spécificité, allant des unités lexicales les plus spécifiques aux moins spécifiques.

Ensuite, ces 4717 spécificités ont fait l'objet d'une deuxième analyse quantitative, à savoir une analyse sémantique quantitative à partir d'une analyse des cooccurrences (*chapitre 5*). A cet effet, nous avons implémenté la monosémie en termes d'homogénéité sémantique, ce qui a permis de quantifier la monosémie d'un mot de base à partir du degré de recoupement formel des cooccurrents de ses cooccurrents. Le degré de recoupement des cooccurrents de deuxième ordre indique effectivement

à quel point les cooccurrents de premier ordre (c'est-à-dire les contextes du mot de base) sont similaires entre eux et donc homogènes sémantiquement. Afin de calculer le degré de recoupement, nous avons élaboré une mesure de recoupement à partir du nombre de cooccurrents de deuxième ordre qui sont partagés par les cooccurrents de premier ordre. Plus le résultat de cette mesure est élevé, plus les cooccurrents de deuxième ordre sont partagés et, par voie de conséquence, le mot de base est plus homogène sémantiquement. Les degrés de monosémie, à l'instar des degrés de spécificité, ont permis de situer les unités lexicales spécifiques sur un continuum sémantique, allant des unités les plus homogènes sémantiquement aux moins homogènes.

La mesure de recoupement élaborée dans le cinquième chapitre, de même que les paramètres de la base de données des cooccurrents de premier et de deuxième ordre, ont ensuite été raffinés et mis au point (*chapitre 6*). Ces mises au point ont abouti à une configuration plus stable de la base de données, ainsi qu'à une mesure de monosémie technique pondérée, en fonction de la technicité (ou de la spécificité) des cooccurrents de deuxième ordre. Enfin, nous avons testé des mesures alternatives pour un échantillon de 50 spécificités représentatives, dans le but de juger la pertinence des facteurs repris dans la mesure de recoupement de base.

- *Résultats des analyses statistiques de base*

Après avoir établi le continuum de spécificité et le continuum sémantique à partir des rangs de spécificité et des rangs de monosémie, nous avons soumis les données quantitatives de la liste des 4717 spécificités du corpus technique à plusieurs analyses statistiques (*chapitre 7*). D'abord, nous avons procédé à une analyse statistique de régression simple, afin d'évaluer l'impact du rang de spécificité sur le rang de monosémie et, plus particulièrement, la corrélation entre le rang de spécificité des 4717 spécificités et leur rang de monosémie. Ensuite, une analyse statistique de régression multiple a permis d'évaluer l'impact combiné de plusieurs variables susceptibles d'influer sur le rang de monosémie d'une unité spécifique, par exemple son rang de spécificité, sa fréquence dans le corpus technique et dans le corpus général, sa longueur, la ou les classe(s) lexicale(s) dont elle fait partie.

Les résultats de l'analyse de régression simple nous ont permis d'infirmer la thèse monosémiste traditionnelle. En effet, ils ont démontré une corrélation négative entre le rang de spécificité et le rang de monosémie des 4717 spécificités du corpus technique. Ainsi, il s'est avéré que les unités lexicales les plus spécifiques du corpus technique ne sont pas les plus monosémiques, mais qu'au contraire, ce sont les plus hétérogènes sémantiquement, par exemple *machine, pièce, tour*. En plus, les unités lexicales les moins spécifiques du corpus technique sont les plus monosémiques (*rationnellement, télédiagnostic, autosurveillance*), à quelques exceptions près,

comme *service, objet, commercial, air, eau*. Notons que pour interpréter correctement les résultats des analyses statistiques, il est indispensable de tenir compte des particularités de la mesure de monosémie sous-jacente et de considérer la monosémie en termes d'homogénéité sémantique et la polysémie et l'homonymie (et le vague) en termes d'hétérogénéité sémantique.

Dans le but de préciser les résultats de la mesure de monosémie de base, nous avons aussi élaboré une mesure de monosémie technique pondérée, enrichie par des informations linguistiques. Cette mesure nous a permis de tenir compte, pendant le calcul du recoupement des cooccurrents des cooccurrents (cc), de la spécificité des cc dans le corpus technique, donc de leur technicité. Toutefois, les résultats pour le rang de monosémie technique sont moins concluants que ceux qui concernent le rang de monosémie de base. Bien qu'elle soit toujours négative, la tendance observée est moins forte. Si on prend en considération la spécificité des cc, les unités lexicales les plus spécifiques se situent toujours du côté des rangs les moins monosémiques (ou les moins homogènes sémantiquement).

Dans l'analyse de régression simple, nous avons été confrontée au problème de l'hétéroscédasticité. L'hétéroscédasticité signifie, rappelons-le, que les estimateurs de la méthode des moindres carrés ne sont pas efficaces et que la droite de régression linéaire n'est pas la meilleure prédiction possible. Par conséquent, le résultat de l'analyse de régression simple, à savoir le pourcentage de variation expliquée, n'est pas fiable. En effet, la corrélation que nous avons observée entre le rang de spécificité et le rang de monosémie ne s'est pas avérée tout à fait linéaire, quelques spécificités se situant très loin de la droite de régression, avec des résidus très importants. Nous avons pu constater que certaines spécificités sont effectivement plus monosémiques que l'on pourrait attendre en fonction de leur rang de spécificité et qu'elles ont des résidus négatifs importants, par exemple *autocalibrage, hydrauliquement, polygonal*. Par contre, d'autres spécificités sont plus polysémiques (ou plus hétérogènes sémantiquement) que l'on pourrait attendre en fonction de leur rang de spécificité, par exemple *service, objet, commercial*, etc. Ces spécificités ont des résidus positifs importants et elles sont plutôt nombreuses, se situant dans la partie supérieure droite de la visualisation de base.

Afin de découvrir l'origine de l'hétéroscédasticité et dans le but de trouver une solution opérationnelle surtout, nous avons d'abord procédé à des analyses exploratoires, en fonction de l'importance des résidus des spécificités et en fonction de la fréquence technique et générale des spécificités. Ensuite, nous avons adopté les solutions techniques les plus courantes, à savoir des transformations logarithmiques et polynomiales, une analyse de régression simple pondérée et une analyse de régression non linéaire. Ces solutions techniques nous ont permis de résoudre le problème de l'hétéroscédasticité et d'aboutir à des pourcentages de variation

expliquée plus élevés et plus fiables. En plus, ces solutions ont confirmé notre hypothèse initiale : les unités lexicales les plus spécifiques du corpus technique ne sont pas les plus monosémiques. Cependant, ces solutions techniques se sont avérées difficiles à interpréter du point de vue linguistique. Etant donné que la visualisation de la régression non linéaire indique que la tendance négative ne s'applique pas à toutes les spécificités, nous avons opté pour la solution d'exclusion d'un sous-ensemble de spécificités.

Le meilleur critère d'exclusion permettant de résoudre l'hétéroscédasticité pour le sous-ensemble restant a été le critère de la fréquence générale. En effet, les mots les plus fréquents dans le corpus de référence de langue générale échappent en partie à la corrélation négative entre le rang de spécificité et le rang de monosémie. Ces 1507 spécificités exclues sont des mots généraux, tels que *service, objet, commercial*, qui se caractérisent par une polysémie à la fois générale et technique, en raison de la percolation de leur polysémie générale dans le corpus technique et en raison du faible recoupement de leurs cc techniques (polysémie technique). Les mots produisent un effet perturbateur par rapport à la tendance générale de corrélation négative et échappent à une prédiction de leur rang de monosémie à partir de leur rang de spécificité, du fait qu'ils sont de toutes façons plutôt polysémiques, quel que soit leur rang de spécificité.

Les 3210 spécificités techniques restantes sont très peu fréquentes ou même absentes du corpus de référence de langue générale. Elles se caractérisent par l'homoscédasticité et par une bonne corrélation linéaire négative entre le rang de spécificité et le rang de monosémie. Il s'ensuit que, parmi les 3210 spécificités, les mots les plus spécifiques du corpus technique sont plutôt hétérogènes sémantiquement, par exemple *usinage, broche, arête*. Par contre, les mots les moins spécifiques sont plutôt homogènes sémantiquement (*adhésif, présentoir*), tant pour le rang de monosémie que pour le rang de monosémie technique. Les résultats de l'analyse de régression simple pour ce sous-ensemble de spécificités techniques conduisent donc également à une remise en question quantitative de la thèse monosémiste. Toutefois, il convient de rappeler la nécessité de recherches supplémentaires à ce sujet. En effet, il faudrait vérifier si et à quel point la « monosémie » des monosémistes correspond exactement à notre mesure de recoupement ou de monosémie, qui implémente la monosémie en termes d'homogénéité sémantique (Cf. 9.2).

Les résultats de l'analyse de régression multiple nous ont permis de confirmer les résultats de l'analyse de régression simple. Ils ont permis également d'apporter des précisions grâce à l'intégration de toutes les variables indépendantes susceptibles d'influencer le rang de monosémie et le rang de monosémie technique. Il s'est avéré que les variables indépendantes significatives expliquent environ 80% de la

variation du rang de monosémie et 75% de la variation du rang de monosémie technique. Nous avons constaté que la variable indépendante la plus significative est la variable du rang de fréquence dans le corpus technique. Comme on pouvait s'y attendre en fait, les spécificités les plus fréquentes dans le corpus technique sont les plus hétérogènes sémantiquement. Les autres variables indépendantes significatives sont le rang de spécificité et, dans l'ordre, le degré de spécificité, la longueur et, finalement, le nombre de classes lexicales. Pour le rang de monosémie comme variable dépendante, le rang de spécificité ou le degré de spécificité se caractérisent par une corrélation négative, ce qui confirme les résultats de l'analyse de régression simple.

- *Résultats des analyses statistiques détaillées*

Finalement, nous avons procédé à des analyses statistiques détaillées, c'est-à-dire à des analyses de régression par classe lexicale et par sous-corpus, ainsi qu'à des analyses pour certaines sous-catégories des 4717 spécificités (*chapitre 8*). Le but de ces analyses statistiques détaillées était de vérifier si les résultats et les conclusions des analyses de base s'appliquent aussi à des sous-ensembles et aux spécificités des sous-corpus. Nous étions particulièrement intéressée par le sous-corpus des normes, parce que ces textes sont censés être prescriptifs et normatifs.

Pour les quatre classes lexicales (adjectifs, adverbes, substantifs et verbes), nous avons observé une corrélation négative entre le rang de spécificité et le rang de monosémie et, respectivement, entre le rang de spécificité et le rang de monosémie technique. Les mots les plus spécifiques d'une classe lexicale sont les plus polysémiques à l'intérieur de cette classe lexicale. Il s'est avéré aussi que la classe lexicale des substantifs illustre le mieux la corrélation négative et corrobore, dès lors, le mieux le pouvoir explicatif du rang de spécificité. Cette constatation renforce à son tour la remise en question de la thèse monosémiste, d'autant plus que les substantifs sont généralement très bien représentés dans les textes et les corpus techniques. Les adverbes illustrent moins bien la corrélation négative. En plus, nous avons constaté que l'explication quantitative des résultats, en termes de spécificités fréquentes dans le corpus général, s'applique aussi aux analyses par classe lexicale. L'explication quantitative s'y accompagne d'une explication linguistique, en termes de caractéristiques syntaxiques et collocationnelles et par opposition des substantifs aux adverbes. Les substantifs se caractérisent par des mécanismes collocationnels et désambiguïsateurs plus puissants, qui se reflètent clairement dans les résultats de la mesure de recoupement et de la mesure de recoupement technique pondérée. Cependant, il convient de signaler que les deux mesures de recoupement ou de monosémie reposent essentiellement sur l'analyse des cooccurrences statistiquement significatives (Cf. 9.2).

Les analyses de régression détaillées par sous-corpus ont conduit à des résultats similaires, aussi bien les analyses de régression simple que les analyses de régression multiple. Elles ont également permis donc de réfuter la thèse monosémiste, notamment dans le sous-corpus des normes et dans celui des manuels. En effet, les unités lexicales les plus spécifiques dans les normes, tout comme dans les manuels, ne sont pas les unités les plus homogènes sémantiquement, mais au contraire les plus hétérogènes. De plus, les normes se sont avérées être les moins sensibles aux effets de la mesure de monosémie technique pondérée. Finalement, nous avons démontré que les unités les plus spécifiques dans les normes (*sécurité, dispositif, risque, exigence*), ne sont pas les unités les plus homogènes sémantiquement dans le corpus entier. Par conséquent, la corrélation négative entre, d'une part, le rang de spécificité dans les normes et, de l'autre, le rang de monosémie dans le corpus entier, nous a permis de corroborer les résultats de notre étude sémantique quantitative et de réfuter une fois de plus la thèse monosémiste.

Toutefois, il est à noter que les résultats des analyses statistiques sont tributaires du corpus technique utilisé ainsi que de la mesure de monosémie. Si les analyses sont conduites sur un autre corpus spécialisé ou si elles s'appuient sur une autre quantification de la monosémie, elles aboutiront probablement à d'autres résultats, qui seront peut-être moins extrêmes.

9.2. PERSPECTIVES

Notre étude a permis d'apporter des réponses quantitatives et linguistiques à des questions sémantiques. A son tour, elle soulève de nouvelles questions, ouvrant la voie à des recherches plus approfondies et à des prolongements, notamment en ce qui concerne les unités polylexicales et la mesure de recoupement.

La poursuite de nos travaux passe inévitablement par les unités polylexicales, étant donné que la plupart des unités lexicales spécifiques d'un corpus technique se situent à ce niveau. Notre approche méthodologique de l'analyse des cooccurrences, qui permet de quantifier l'analyse sémantique, est facilement transposable aux unités polylexicales. D'abord, pour l'identification des unités polylexicales, on pourra soit recourir à des logiciels d'extraction automatique d'unités terminologiques, soit considérer les spécificités actuelles (mots simples) comme des mots de base et identifier leurs cooccurrents statistiquement très pertinents. Ensuite, les unités polylexicales relevées pourront être considérées comme de nouvelles unités de base, dont on analysera les cooccurrents de deuxième ordre. Cette analyse reviendra dès lors à l'analyse des cooccurrents de troisième ordre par rapport aux spécificités qui sont à l'origine de ces unités polylexicales. De telle façon, on pourra parfaitement attribuer un degré d'homogénéité sémantique aux unités polylexicales afin de les

classer par ordre décroissant. On recensera probablement peu d'unités polylexicales hétérogènes sémantiquement, étant donné que le deuxième composant (ou le cooccurrent de premier ordre pertinent) entraîne déjà une certaine désambiguïsation de l'unité polylexicale (Cf. les mots composés des analyses de régression détaillées du chapitre 8).

Toutefois, les données sémantiques quantitatives des unités polylexicales, c'est-à-dire leurs rangs de monosémie (ou d'homogénéité sémantique) et de monosémie technique, ne pourront pas faire l'objet d'une analyse statistique de régression simple telle que nous l'avons effectuée dans la présente étude. Comme nous l'avons déjà évoqué dans la section sur les restrictions (Cf. chapitre 1), il est techniquement très difficile de déterminer le degré de spécificité des unités polylexicales par l'intermédiaire de la méthode des mots-clés. D'une part, les unités polylexicales du corpus technique sont tellement spécifiques et tellement techniques, qu'elles sont majoritairement absentes du corpus de référence de langue générale. D'autre part, le fait d'être constitué de plusieurs unités lexicales simples complique les calculs propres à la méthode des mots-clés. A ce sujet, nous nous proposons également d'approfondir la méthodologie de l'analyse des spécificités dans le but de déterminer la spécificité ou la technicité des unités (poly)lexicales de façon alternative. D'ailleurs, il s'est avéré que la méthode des mots-clés et, plus particulièrement la mesure statistique du rapport de vraisemblance (LLR), est légèrement sensible aux fréquences élevées dans le corpus d'analyse. Toutefois, la méthode des mots-clés s'est révélée très utile pour classer, par ordre décroissant de spécificité, les unités lexicales simples qui sont spécifiques ou représentatives de notre corpus technique.

Notre étude quantitative a démontré l'hétérogénéité sémantique de nombreuses unités lexicales simples, spécifiques du corpus technique. Leur hétérogénéité sémantique s'explique en partie par le fait que ces unités lexicales, telles que *machine*, entrent souvent dans la composition d'unités polylexicales, comme *machine à usiner* ou *machine à rectifier*. Ainsi, il y aura probablement une corrélation entre, d'une part, une unité simple plus hétérogène sémantiquement et, de l'autre, le nombre plus élevé d'unités polylexicales pertinentes qui la contiennent et qui sont homogènes sémantiquement. A titre de comparaison, il serait intéressant d'étudier également la sémantique des combinaisons libres, non idiomatiques.

Nous aimerions aussi formuler quelques possibles améliorations de notre mesure de recoupement. D'abord, il serait intéressant de vérifier si et à quel point la « monosémie » de l'approche traditionnelle correspond au degré ou au rang de monosémie « monosémique » calculé par notre mesure de recoupement. En plus, la mesure de recoupement actuelle pourra être enrichie si on y intègre plus d'informations linguistiques. En effet, lors de la génération de la base de données

des mots de base (spécificités) et des cooccurrents de premier et de deuxième ordre, on pourra intégrer soit le code Cordial, soit un nouveau code qui indique la catégorie grammaticale, à partir du code Cordial. Ces informations permettront de privilégier certains cooccurrents du mot de base en fonction de leur catégorie grammaticale, éventuellement sous forme de pondération. Bien entendu, cet enrichissement linguistique risque de compliquer la formule de la mesure de recoupement et les calculs, puisqu'il faut qu'on adapte la formule en fonction de la catégorie grammaticale du mot de base. En ce qui concerne la base de données des cooccurrents de premier et de deuxième ordre, nous envisageons en outre de tenir compte des frontières de documents, dans le but d'affiner au mieux les analyses des cooccurrences. A cet effet, le corpus technique requiert des opérations de nettoyage supplémentaires, en particulier l'insertion de délimiteurs indiquant la fin des documents.

Dans notre étude quantitative, nous avons implémenté l'analyse sémantique en termes d'homogénéité sémantique et d'hétérogénéité sémantique. En effet, cette reformulation opérationnelle est incontournable dans une approche quantitative, qui aboutit à une analyse statistique de régression. Cependant, afin d'affiner les résultats et les interprétations, nous projetons de compléter la mesure de monosémie élaborée dans notre thèse par des analyses statistiques multivariées de regroupement (*cluster analysis*). Celles-ci permettraient de regrouper les cooccurrents (ou c) d'un mot de base (spécificité) à partir des cc qu'ils partagent. Les analyses de regroupement conduiraient peut-être à mieux comprendre encore le phénomène de l'hétérogénéité sémantique et à opérer des distinctions sémantiques plus fines. Comme nous l'avons déjà évoqué à plusieurs reprises, notre approche sémantique quantitative ne permet pas (encore) de distinguer entre la polysémie et le sens vague. Bien sûr, cela tient principalement à l'implémentation opérationnelle en termes d'hétérogénéité sémantique, mais nous aimerions aussi invoquer que les critères traditionnels, notamment entre la polysémie et le vague, sont particulièrement vagues.

Nous avons démontré que l'approche quantitative comporte de nombreux avantages. D'abord, elle permet de procéder à l'analyse sémantique simultanée de plusieurs milliers d'unités lexicales. Ensuite, les données quantitatives qui en résultent se prêtent à des analyses statistiques, qui conduisent à des résultats objectifs. Enfin, les approches méthodologiques élaborées dans notre étude, à savoir la double analyse quantitative et les analyses statistiques de régression, pourront facilement être appliquées à d'autres corpus spécialisés, relevant d'autres domaines techniques ou scientifiques, tels que l'électronique ou l'informatique. On pourrait même envisager une analyse sémantique quantitative, par le biais d'une analyse des cooccurrences, pour un corpus de langue générale.

En guise de conclusion, nous aimerions rappeler que l'analyse élaborée ici se prête certainement à la mise au point de la mesure de recoupement actuelle et à l'application d'autres mesures qui intégreraient d'autres facteurs. En plus, notre analyse sémantique quantitative mérite d'être appliquée à d'autres unités, en particulier aux unités polylexicales de notre corpus technique, ainsi qu'à d'autres corpus d'analyse, par exemple à un corpus de langue générale. Ces analyses complémentaires se situent clairement dans le prolongement de notre étude, puisque notre thèse de doctorat ne constitue qu'une étape dans l'étude sémantique quantitative de la langue spécialisée.

Bibliographie

Adelstein, A. & M.T. Cabré

2002 The specificity of units with specialized meaning : polysemy as explanatory factor. *DELTA* 18 : 1-25.

Arntz, R. & H. Picht

1989 *Einführung in die Terminologiearbeit*. Hildesheim : Georg Olms Verlag.

Audibert, L.

2001 LoX : outil polyvalent pour l'exploration de corpus annotés. *Actes de RECITAL (TALN) 2001* : 411-419.

2002 Etude des critères de désambiguïsation sémantique automatique : présentation et premiers résultats sur les cooccurrences. *Actes de RECITAL (TALN) 2002* : 415-424.

2003 Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. *Actes de TALN 2003* : 35-44.

Beaudoin, V.

2000 Statistique textuelle : une approche empirique du sens à base d'analyse distributionnelle. *Texto !* http://www.revue-texto.net/Inedits/Beaudouin_Statistique.html

Béjoint, H. & P. Thoiron

2000 Le sens des termes. In H. Béjoint & P. Thoiron (Eds.), *Le sens en terminologie* 5-19. Lyon : Presses universitaires de Lyon.

2002 Schéma définitionnel, définition et traitement lexicographique des termes. *Cahiers de lexicologie* 80(1) : 121-134.

Béjoint, H.

1989 A propos de la monosémie en terminologie. *Meta* 34(3) : 405-411.

Berber Sardinha, A.

1996 Review : WordSmith Tools. *Computers & Texts* 12 : 19-21.

1999a Word sets, keywords and text contents : an investigation of text topic on the computer. *DELTA* 15(1) : 141-149.

1999b Using KeyWords in text analysis : practical aspects. *DIRECT Papers* 42 : 1-8.

Bergenholtz, H. & U. Kaufmann

1997 Terminography and lexicography. A critical survey of dictionaries from a single specialised field. *Hermes* 18 : 91-125.

Bertels, A.

2005 A la découverte de la polysémie des spécificités du français technique. *Actes de RECITAL (TALN) 2005* : 575-584.

Bertels, A., D. Speelman & D. Geeraerts

2006 Analyse quantitative et statistique de la sémantique dans un corpus technique. *Actes de TALN 2006* : 73-82.

Bianchi, C.

2001 La flexibilité sémantique : une approche critique. *Langue française* 129 : 91-109.

Biber, D.

1995 *Dimensions of register variation. A cross-linguistic comparison.* Cambridge : Cambrigde University Press.

Biber, D., S. Conrad & R. Reppen

1998 *Corpus linguistics. Investigating language structure and use.* Cambridge : Cambrigde University Press.

Bourigault, D. & C. Frérot

2005 Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. *Actes de TALN 2005* : 373-382.

Bourigault, D. & M. Slodzian

1999 Pour une terminologie textuelle. *Terminologies Nouvelles* 19 : 29-32.

Bourigault, D.

1994 *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes.* Thèse en informatique linguistique, Ecole des hautes Etudes en Sciences Sociales, Paris.

Bourigault, D., C. Jacquemin & M.-C. L'Homme

2001 *Recent advances in computational terminology*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bouveret, M.

1998 Approche de la dénomination en langue spécialisée. *Meta* 43(3) : 1-18.

Bowker, L. & J. Pearson

2002 *Working with specialized language. A practical guide to using corpora*. London : Routledge.

Brunet, E.

2000 Qui lemmatise dilemme attise. *Lexicometrica* 2. http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2/brunet2000.PDF

2002 *Hyperbase©. Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels. Manuel de référence*. Université de Nice. http://perso.orange.fr/hyperbas/manuel.pdf

Cabré, M.T.

1991 Terminologie ou terminologies ? Spécialité linguistique ou domaine interdisciplinaire ? *Meta* 36(1) : 55-63.

1998 *La terminologie. Théorie, méthode et applications*. Ottawa : Les Presses de l'Université.

2000a Terminologie et linguistique : la théorie des portes. *Terminologies nouvelles* 21 : 10-15.

2000b Sur la représentation mentale des concepts : bases pour une tentative de modélisation. In H. Béjoint & P. Thoiron (Eds.), *Le sens en terminologie* 20-39. Lyon : Presses universitaires de Lyon.

Cadiot, P. & B. Habert

1997 Aux sources de la polysémie nominale. *Langue française* 113 : 3-11.

Candel, D.

1994 *Français scientifique et technique et dictionnaire de langue*. Paris : Didier Erudition.

Cherfi, H. & Y. Toussaint

2002 Adéquation d'indices statistiques à l'interprétation de règles d'association. *Actes de JADT 2002*. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/cherfi_toussaint.pdf

Chetouani, L. & S. Heiden

2000 Sémantique des noms propres. Méthode des cooccurrences. *Actes de JADT 2000* : 29-32.

Chung, Y.M. & J.Y. Lee

2001 A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American society for information science and technology* 52(4) : 283-296.

Church, K.W. & P. Hanks

1990 Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1) : 22-29.

Condamines, A. & J. Rebeyrolle

1997 Point de vue en langue spécialisée. *Meta* 42(1) : 174-184.

Condamines, A.

1999 Approche sémasiologique pour la constitution de Bases de Connaissances Terminologiques. In V. Delavigne & M. Bouveret (Eds.), *Sémantique des termes spécialisés* 101-117. Rouen : Publications de l'Université de Rouen.

Conrad, S. & D. Biber

2001 *Variation in English. Multi-dimensional studies*. Harlow : Pearson Education Limited.

Cruse, D.A.

1986 *Lexical semantics*. Cambridge : Cambridge University Press.

2000 *Meaning in language. An introduction to semantics and pragmatics*. Oxford : Oxford University Press.

Cuyckens, H. & B. Zawada

1997 *Polysemy in cognitive linguistics. Selected papers from the Fifth International Cognitive Linguistics Conference*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

De Marneffe, M.-C. & P. Dupont

2004 Comparative study of statistical word sense discrimination techniques. *Actes de JADT 2004* : 270-281.

De Vogüé, S. & D. Paillard

1997 Identité lexicale et hétérogénéité de la variation co-textuelle. Le cas de *suivre*. In C. Guimier (Ed.), *Co-texte et calcul du sens* 41-61. Caen : Presses Universitaires de Caen.

Delavigne, V. & M. Bouveret

1999 *Sémantique des termes spécialisés*. Rouen : Publications de l'Université de Rouen.

Delavigne, V.

2003 Quand le terme entre en vulgarisation. *Actes de Terminologie et Intelligence Artificielle TIA 2003* : 80-91.

Denhière, G. & B. Lemaire

2003 Modélisation des effets contextuels par l'analyse de la sémantique latente. *Actes des Deuxièmes Journées d'étude en Psychologie Ergonomique (EPIQUE 2003).* http://www.upmf-grenoble.fr/sciedu/blemaire/epique03.pdf

Dorow, B & D.Widdows

2003 Discovering corpus-specific word senses. *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics, Budapest, Hungary* : 79-82.

Drouin, P.

2003a Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1) : 99-117.

2003b Acquisition des termes simples fondée sur les pivots lexicaux spécialisés. *Actes de Terminologie et Intelligence Artificielle TIA 2003* : 183-186.

2004 Spécificités lexicales et acquisition de la terminologie. *Actes de JADT 2004* : 345-352.

Dunning, T.

1993 Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1) : 61-74.

Dury, P.

1999 Les variations sémantiques en terminologie : étude diachronique et comparative appliquée à l'écologie. In V. Delavigne & M. Bouveret (Eds.), *Sémantique des termes spécialisés* 17-32. Rouen : Publications de l'Université de Rouen.

Ellman, J., I. Klincke & J. Tait

2000 Word Sense Disambiguation by information filtering and extraction. *Computers and the Humanities* 30(1-2) : 127-134.

Eriksen, L.

2002 Die Polysemie in der Allgemeinsprache und in der juristischen Fachsprache. Oder : Zur Terminologie der 'Sache' im Deutschen. *Hermes* 28 : 211-222.

Evert, S. & B. Krenn

2001 Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France* : 188-195.

2003 Computational approaches to collocations. Introductory course at the *European Summer School on Logic, Language, and Information (ESSLLI 2003), Vienna*. http://www.collocations.de/EK/index.html

Evert, S. & H. Kermes

2003 Experiments on candidate data for collocation extraction. *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics, Budapest, Hungary* : 83-86.

Evert, S.

2002 Special topic session on the mathematical properties of association measures. Presentation at the *Workshop on Computational Approaches to Collocations. Vienna*. http://www.collocations.de/EK/index.html

Fabre, C., B. Habert & D. Labbé

1997 La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques* 13 : 15-30.

Ferrari, L.

2002 Un caso de polisemia en el discurso jurídico ? *Terminology* 8(2) : 221-244.

Ferret, O.

2004 Découvrir des sens de mots à partir d'un réseau de cooccurrences lexicales. *Actes de TALN 2004*. http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/Ferret.pdf

Firth, J.R.

1957 Modes of Meaning. *Papers in Linguistics* : 190-215.

François, J.

1997 Le cadrage cognitif des prédications de contact dans un corpus de déclarations d'accidents de la route. Effets du contexte et du co-texte. In C. Guimier (Ed.), *Co-texte et calcul du sens* 73-88. Caen : Presses Universitaires de Caen.

Friel, C.M.

2005 *Advanced Statistics II. Weighted least-squares regression*. Course CJ 789. Sam Houston State University. Texas. http://www.shsu.edu/~icc_cmf/cj_789/weighted LeastSquares2.doc

Fuchs, C.

1994 *Paraphrase et énonciation*. Paris : Ophrys.

1996 *Les ambiguïtés du français*. Paris : Ophrys.

Fukushige, Y. & N. Noguchi

2000 Statistical and linguistic approaches to automatic term recognition : NTCIR experiments at Matsushita. *Terminology* 6(2) : 257-286.

Gale, W., K. Church & D. Yarowsky

1993 A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26 : 415-439.

Gambier, Y.

1991 Travail et vocabulaire spécialisés : prolégomènes à une socio-terminologie. *Meta* 36(1) : 8-15.

Gaudin, F.

1993 *Pour une socioterminologie. Des problèmes sémantiques aux pratiques institutionnelles*. Rouen : Publications de l'Université de Rouen.

1995a Dire les sciences et décrire les sens : entre vulgarisation et lexicographie, le cas des dictionnaires de sciences. *Traduction, terminologie, rédaction TTR* 8(2) : 11-27.

1995b Champs, clôtures et domaines : des langues de spécialités à la culture scientifique. *Meta* 40(2) : 229-237.

2003 *Socioterminologie* : *une approche sociolinguistique de la terminologie*. Bruxelles : Duculot.

2005 Point de vue d'un socioterminologue. *Actes de Terminologie et Intelligence Artificielle TIA 2005*. http://www.loria.fr/~yannick/TIA2005/doc/gaudin.pdf

Gaume, B., N. Hathout & P. Muller

2004 Désambiguïsation par proximité structurelle. *Actes de TALN 2004* : 205-214.

Geeraerts, D.

1986 *Woordbetekenis. Een overzicht van de lexicale semantiek*. Leuven/Amersfoort : Acco.

1989 *Wat er in een woord zit. Facetten van de lexicale semantiek*. Leuven : Peeters.

1993 Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics* 4(3) : 223-272.

2002 The theoretical and descriptive development of lexical semantics. In L. Behrens & D. Zaefferer (Eds.), *The lexicon in focus. Competition and convergence in current lexicology* 23-42. Frankfurt : Peter Lang Verlag.

Gémar, J.-C.

1991 Terminologie, langue et discours juridiques. Sens et signification du langage du droit. *Meta* 36(1) : 275-283.

Gentilhomme, Y.

2001 Peut-on parler de culture technoscientifique ? *Cahiers de lexicologie* 78 : 107-115.

Gévaudan, P.

1997 La polysémie verticale. Hypothèses, analyses et interprétations. *Philologie im Netz* : *PhiN* 2/1997 : 1-22.

Grefenstette, G.

1994 Corpus-derived first, second and third-order word affinities. In W. Martin, W. Meijs, e.a. (Eds.), *Proceedings of Euralex '94. International Congress on Lexicography, Amsterdam* : 279-290.

Guespin, L.

1995 La circulation terminologique et les rapports entre science, technique et production. *Meta* 40(2) : 206-215.

Guilbert, L.

1973 La spécificité du terme scientifique et technique. *Langue française* 17 : 5-17.

Habert, B., A. Nazarenko & A. Salem

1997 *Les linguistiques de corpus*. Paris : Armand Colin/Masson.

Habert, B., G. Illouz & H. Folch

2004 Dégrouper les sens : pourquoi ? comment ? *Actes de JADT 2004* : 565-576.

2005 Des décalages de distribution aux divergences d'acception. In A. Condamines (Ed.), *Sémantique et corpus* 277-318. Paris : Lavoisier/ Hermes-Science.

Hahn, W. Von

1983 *Fachkommunikation. Entwicklung, linguistische Konzepte, betriebliche Beispiele*. Berlin/New York : Mouton de Gruyter.

1998 Vagheit bei der Verwendung von Fachsprachen. In L. Hoffmann, H. Kalverkämper & H.E. Wiegand (Eds.), *Fachsprachen. Ein Internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft*. Band 1, 378-382. Berlin/New York : Mouton de Gruyter.

Hanks, P.

2000 Do word meanings exist ? *Computers and the Humanities* 30(1-2) : 205-215.

Hausmann, F.J.

1979 Un dictionnaire des collocations est-il possible ? *Travaux de linguistique et de littérature* 17(1) : 187-195.

Heiden, S.

2004 Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex. *Actes de JADT 2004* : 577-588.

Higgins, D.

2004 Which statistics reflect semantics ? Rethinking synonymy and word similarity. *Proceedings of the International Conference on Linguistic Evidence* : 61-65.

Hisamitsu, T. & Y. Niwa

2001 Topic-word selection based on combinatorial probability. *Proceedings of NLPRS 2001* : 289-296.

2002 A measure of term representativeness based on the number of co-occurring salient words. *Proceedings of COLING 2002* : 325-331.

Hisamitsu, T., Y. Niwa, S. Nishioka, H. Sakurai, O. Imaichi, M. Iwayama & A. Takano

2000 Extracting terms by a combination of term frequency and a measure of term representativeness. *Terminology* 6(2) : 211-232.

Hisamitsu, T., Y. Niwa & J. Tsujii

2000 A method of measuring term representativeness. Baseline method using co-occurrence distribution. *Proceedings of COLING 2000* : 320-326.

Hofmann, T.

1999 Probabilistic Latent Semantic Analysis. *Proceedings Fifteenth Conference on Uncertainty in Artificial Intelligence UAI, Stockholm, Sweden*. www.cs.brown.edu/~th/papers/Hofmann-UAI99.ps

Huet, S., A. Bouvier, M.-A. Poursat & E. Jolivet

2004 *Statistical tools for non-linear regression. A practical guide with S-Plus and R examples*. New York/Berlin/Heidelberg : Springer-Verlag.

Humbley, J.

1997 Is terminology specialized lexicography ? The experience of French-speaking countries. *Hermes* 18 : 13-31.

Huot, H.

1996 *Revue française de linguistique appliquée. Dossier* : *Corpus* : *de leur constitution à leur exploitation.* Volume 1(2). Amsterdam : Editions De Werelt.

Ide, N. & J. Véronis

1998 Introduction to the special issue on Word Sense Disambiguation : the state of the art. *Computational Linguistics* 24(1) : 1-40.

Ide, N.

2000 Cross-lingual sense determination : Can it work ? *Computers and the Humanities* 30(1-2) : 223-234.

ISO1087-1

1990 *Travaux terminologiques - Vocabulaire - Partie 1 Théorie et application.* Genève : ISO.

Jacques, M.P.

2003 *Approche en discours de la réduction des termes complexes dans les textes spécialisés*. Thèse de doctorat. Université de Toulouse-Le Mirail.

Jacquet, G. & F. Venant

2005 Construction automatique de classes de sélection distributionnelle. *Actes de TALN 2005* : 303-312.

Ji, H., S. Ploux & E. Wehrli

2003 Lexical knowledge representation with contexonyms. *Proceedings of the 9th MT summit* : 194-201.

Kageura, K.

2002 *The dynamics of terminology. A descriptive theory of term formation and terminological growth*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

Karov, Y. & S. Edelman

1998 Similarity-based Word Sense Disambiguation. *Computational Linguistics* 24(1) : 41-59.

Kayser, D.

1987 Une sémantique qui n'a pas de sens. *Langages* 87 : 33-45.

1989 Réponse à Kleiber et Riegel. *Linguisticae Investigationes* XIII (2) : 419-422.

1995 Terme et dénotation. *La banque des mots* numéro spécial 7 : 19-34.

Kilgarriff, A. & M. Palmer

2000 Introduction to the special issue on SENSEVAL. *Computers and the Humanities* 30 (1-2) : 1-13.

Kleiber, G. & M. Riegel

1989 Une sémantique qui n'a pas de sens n'a vraiment pas de sens. *Linguisticae Investigationes* XIII (2) : 405-417.

1991 Sens lexical et interprétations référentielles. Un écho à la réponse de D. Kayser. *Linguisticae Investigationes* XV (1) : 181-201.

Kleiber, G.

1990 *La sémantique du prototype. Catégories et sens lexical*. Paris : PUF

1994 *Nominales. Essais de sémantique référentielle*. Paris : Armand Colin.

1996 Noms propres et noms communs : un problème de dénomination. *Meta* 41(4) : 567-589.

1997 Quand le contexte va, tout va et … inversement. In C. Guimier (Ed.), *Co-texte et calcul du sens* 11-29. Caen : Presses Universitaires de Caen.

1999 *Problèmes de sémantique. La polysémie en questions*. Lille : Presses Universitaires du Septentrion.

2002 De la polysémie en général à la polysémie prototypique en particulier. *Cahiers de lexicologie* 80 : 89-103.

2004 Y a-t-il des micro-sens ? *Communication présentée aux Journées d'hommage en souvenir de H. Geckeler*. Université de Münster.

Kleiber, G., C. Schnedecker & J.-E. Tyvaert

1997 *La continuïté référentielle*. Metz : Université de Metz.

Klepousniotou, E.

2002 The processing of lexical ambiguity : homonymy and polysemy in the mental lexicon. *Brain and Language* 81 : 205-223.

Kocourek, R.

1991a *La langue française de la technique et de la science*. Wiesbaden : Brandstetter Verlag.

1991b Textes et termes. *Meta* 36(1) : 71-76.

L'Homme, M.C.

1995 Définition d'une méthode de recensement et de codage des verbes en langue technique : applications en traduction. *Traduction, terminologie, rédaction TTR* 8(2) : 67-88.

1997 Méthode d'accès informatisé aux combinaisons lexicales en langue technique. *Meta* 42(1) : 15-23.

2000 Les enseignements d'un mot polysémique sur les modèles de la terminologie. *Cahiers de Grammaire* 25 : 71-91.

2001 Combinaisons lexicales spécialisées. Regroupement des mots clés par classes conceptuelles. In B. Daille & G. Williams (Eds.), *Journées d'étude de l'ATALA. La collocation. Rapport de recherche* 19-22. Nantes : Institut de recherche en informatique de Nantes.

Labbé, C. & D. Labbé

2001 Que mesure la spécificité du vocabulaire ? *Lexicometrica* 3. http://www.cavi.univ-paris3.fr/lexicometrica/article/numero3/specificite2001.PDF

Lafon, P.

1984 *Dépouillements et statistiques en lexicométrie*. Genève/Paris : Slatkine/ Champion.

Lamalle, C., W. Martinez, S. Fleury & A. Salem et al.

2003 *Outils de statistique textuelle. Manuel d'utilisation de Lexico3*. Paris : Université de Paris3.

Lamiroy, B. & M. Charolles

2004 Des adverbes aux connecteurs : le cas de *mais*, *seulement*, *simplement*, *heureusement* et *malheureusement*. *Travaux de linguistique* 49 : 57-79.

2005 Utilisation de corpus pour l'évaluation d'hypothèses linguistiques : étude de *autrement*. In A. Condamines (Ed.), *Sémantique et corpus* 109-145. Paris : Lavoisier/Hermes-Science.

Lamiroy, B.

1998 Prédication et auxiliaires. In M. Forsgren, K. Jonasson & H. Kronning (Eds.), *Prédication, Assertion, Information* 285-299. Uppsala : Acta Universitatis Uppsaliensis.

Landauer, T.K.

2002 Applications of Latent Semantic Analysis. *24th Annual Meeting of the Cognitive Science Society, August 9th 2002*. http://www.knowledge-technologies.com/papers/Cog-Sci-03.pdf

Landauer, T.K., P.W. Foltz & D. Laham

1998 Introduction to Latent Semantic Analysis. *Discourse Processes* 25 : 259-284.

Lapata, M.

2002 The disambiguation of nominalizations. *Computational Linguistics* 28(3) : 357-388.

Lebart, L. & A. Salem

1994 *Statistique textuelle*. Paris : Dunod.

Legendre, L. & P. Legendre

1983 *Numerical ecology*. Amsterdam : Elsevier.

Lemay, C., M.C. L'Homme & P. Drouin

2005 Two methods for extracting specific single-word terms from specialized corpora. Experimentation and evaluation. *International Journal of Corpus Linguistics* 10(2) : 227-255.

Lerat, P.

1995a Terme, mot, vocable. *La banque des mots* numéro spécial 7 : 5-9.

1995b *Les langues spécialisées*. Paris : PUF.

Lethuillier, J.

1991 Combinatoire, terminologies et textes. *Meta* 36(1) : 92-100.

Lin, D.

2000 Word Sense Disambiguation with a similarity-smoothed case library. *Computers and the Humanities* 30(1-2) : 147-152.

Loupy, C. de

2002 Evaluation des taux de synonymie et de polysémie dans un texte. *Actes de RECITAL (TALN) 2002* : 225-234.

Loupy, C. de, M. El-Bèze & P.-F. Marteau

2000 Using semantic classification trees for WSD. *Computers and the Humanities* 30(1-2) : 187-192.

Manguin, J.-L., J. François & B. Victorri

2005 Polysémie adjectivale et rection nominale : quand *gros* et *gras* sont synonymes. In J. François (Ed.), *L'adjectif en français et à travers les langues* 521-540. Caen : Presses Universitaires de Caen.

Manning, C. & H. Schütze

2002 *Foundations of Statistical Natural Language Processing*. Cambridge (MA) : MIT Press.

Martinez, W. & M. Zimina

2002 Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. *Actes de JADT 2002* : 495-506.

Martinez, W.

2000 Mise en évidence de rapports synonymiques par la méthode des cooccurrences. *Actes de JADT 2000* : 78-84.

Mel'čuk, I.A., A. Clas & A. Polguère

1995 *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Editions Duculot.

Meyer, I. & K. Mackintosh

2000 L'étirement du sens terminologique : aperçu du phénomène de la déterminologisation. In H. Béjoint & P. Thoiron (Eds.), *Le sens en terminologie* 198-217. Lyon : Presses universitaires de Lyon.

Müller, C.

1992a *Initiation aux méthodes de la statistique linguistique* (réimp. de l'édition de 1968). Paris : Champion.

1992b *Principes et méthodes de statistique lexicale* (réimp. de l'édition de 1977). Paris : Champion.

Nakagawa, H.

2000 Automatic term recognition based on statistics of compound nouns. *Terminology* 6(2) : 195-210.

Nerlich, B., Z. Todd, V. Herman & D. Clarke

2003 *Polysemy. Flexible patterns of meaning in mind and language*. Berlin/New York : Mouton de Gruyter.

Nielsen, F.

2002 *Linear Regression Models. Module 12 Heteroscedasticity*. Course Soci209. University of North Carolina. http://www.unc.edu/~nielsen/soci209/m12/m12.htm

Normand, S.

1999 Construction du sens dans un échange professionnel lié à la dégustation. In V. Delavigne & M. Bouveret (Eds.), *Sémantique des termes spécialisés* 119-126. Rouen : Publications de l'Université de Rouen.

Nyckees, V.

1998 *La sémantique*. Paris : Belin.

Oguy, A.

1998 Probleme der experimentellen Erforschung der Wortbedeutung. Überblick über Polysemieuntersuchungen. *Sprachwissenschaft* 23(1) : 113-140.

1999 Approximativ-quantitative Charakteristika der Polysemie. *Sprachwissenschaft* 24(1) : 75-103.

Opitz, K.

1990 The Technical Dictionary for the Expert. In F.-J. Hausmann, O. Reichmann, E. Wiegand et L. Zgusta (Eds.), *Wörterbücher/ Dictionaries/ Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie* (2) 1505-1512. Berlin/New York : Mouton de Gruyter.

Paillard, M.

1997 Co-texte, collocations, lexique. In C. Guimier (Ed.), *Co-texte et calcul du sens* 63-71. Caen : Presses Universitaires de Caen.

Pavel, S.

1991 Changement sémantique et terminologie. *Meta* 36(1) : 41-48.

Pearson, J.

1998 *Terms in context*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

Péroz, P.

2002 Le mot *clé*. Variations sémantiques et régularité des fonctionnements. *Langue française* 133 : 42-53.

Pezik, P.

2005 You shall know a word by the company it keeps. A comparative study of co-occurrence statistics. Paper presented at *PALC 2005, Practical applications in language and computers, Lodz, Poland*.

Phal, A.

1971 *Vocabulaire général d'orientation scientifique (V.G.O.S.). Part du lexique commun dans l'expression scientifique*. Paris : CREDIF/Didier.

Picoche, J.

1986 *Structures sémantiques du lexique français*. Paris : Nathan.

Piot, M.

1996 Propriétés et définition des conjonctions de subordination, de coordination, et adverbes conjonctifs en français. *Leuvense bijdragen* 84(3) : 329-348.

Poibeau, T.

2004 Pré-analyse de corpus. *Actes de JADT 2004* : 897-903.

Portelance, C.

1991 Fondements linguistiques de la terminologie. *Meta* 36(1) : 64-70.

Pustejovski, J. & B. Boguraev

1996 *Lexical semantics. The problem of polysemy*. Oxford : Clarendon Press.

Pustejovski, J.

1995 *The generative lexicon*. Cambridge/Massachusetts : MIT Press.

R Development Core Team

2004 *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org

Rastier, F.

1994 *Sémantique pour l'analyse. De la linguistique à l'informatique*. Paris : Masson.

1995 Le terme : entre ontologie et linguistique. *La banque des mots* numéro spécial 7 : 35-65.

1996 *Sémantique interprétative*. Paris : PUF.

2003 De la signification au sens. Pour une sémiotique sans ontologie. *Texto !* http://www.revue-texto.net/Inedits/Rastier/Rastier_Semiotique-ontologie.html

2006 De la signification lexicale au sens textuel : éléments pour une approche unifiée. *Texto !* http://www.revue-texto.net/Inedits/Rastier/Rastier_Signification-lexicale.html

Ravin, Y. & C. Leacock

2000 *Polysemy. Theoretical and computational approaches.* Oxford : Oxford University Press.

Rayson, P. & R. Garside

2000 Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000)* : 1-6.

Récanati, F.

1997 La polysémie contre le fixisme. *Langue française* 113 : 107-123.

Resche, C.

1999 Equivocal economic terms or terminology revisited. *Meta* 44(4) : 617-632.

Resnik, P. & D. Yarowsky

1997 A perspective on word sense disambiguation methods and their evaluation. *Proceedings of SIGLEX '97, Washington DC* : 79-86.

2000 Distinguishing systems and distinguishing senses : new evaluation methods for word sense disambiguation. *Natural Language Engineering* 5(3) : 113-133.

Riggs, F.W.

1989 Terminology and lexicography : their complementarity. *International Journal of Lexicography* 2(2) : 89-110.

Roche, M., T. Heitz, O. Matte-Tailliez & Y. Kodratoff

2004 EXIT : un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. *Actes de JADT 2004* : 948-956.

Ross, S.

1994 *A first course in probability*. New York : Macmillan College Publishing Company.

Rossignol, M. & P. Sébillot

2002 Automatic generation of sets of keywords for theme characterization and detection. *Actes de JADT 2002*. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/ jadt2002/PDF-2002/rossignol_sebillot.pdf

Ruhl, C.

1989 *On monosemy. A study in linguistic semantics*. Albany, N.Y. : State University of New York Press

Sager, J.

2000 Pour une approche fonctionnelle de la terminologie. In H. Béjoint & P. Thoiron (Eds.), *Le sens en terminologie* 40-60. Lyon : Presses universitaires de Lyon.

Schütze, H.

1998 Automatic Word Sense Discrimination. *Computational Linguistics* 24(1) : 97-123.

Scott, M.

1999 *WordSmith Tools. Version 3*. Oxford : Oxford University Press.

Sébillot, P.

1998 Acquérir des informations sémantiques à partir de corpus. *Workshop des ateliers en morphologie, Colex, La structure du lexique*, Nantes : 173-181.

Segond, F.

2000 Framework and results for French. *Computers and the Humanities* 30 (1-2) : 49-60.

Segond, F., E. Aimelet et al.

2000 Dictionary-driven semantic look-up. *Computers and the Humanities* 30 (1-2) : 193-197.

Sinclair, J.

1991 *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Slodzian, M.

1995 Comment revisiter la doctrine terminologique aujourd'hui ? *La banque des mots* numéro spécial 7 : 11-18.

2000 L'émergence d'une terminologie textuelle et le retour du sens. In H. Béjoint & P. Thoiron (Eds.), *Le sens en terminologie* 61-85. Lyon : Presses universitaires de Lyon.

Speelman, D.

1997 *Abundantia verborum : a computer tool for carrying out corpus-based linguistic case studies.* PhD Thesis. K.U.Leuven.

2005 *Methods of Corpus Linguistics*. Cours F0TU1A. KULeuven.

Stevenson, M. & Y. Wilks

2001 The interaction of knowledge sources in Word Sense Disambiguation. *Computational Linguistics* 27(3) : 321-349.

Stubbs, M.

1995 Collocations and semantic profiles : on the cause of the trouble with quantitative studies. *Functions of language* 2(1) : 23-55.

Suderman, K.

2000 Simple Word Sense Discrimination. *Computers and the Humanities* 30 (1-2) : 165-170.

Temmerman, R.

1997 Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology. *Hermes* 18 : 51-90.

2000a *Towards new ways of terminology description. The sociocognitive approach*. Amsterdam/Philadelphia : John Benjamins Publishing Company.

2000b Une théorie réaliste de la terminologie : le sociocognitivisme. *Terminologies nouvelles* 21 : 58-64.

Tschätsch, H.

1997 *Verspaningstechniek. Technieken en machines*. Traduction néerlandaise. Academic ServiceWetenschap en Techniek.

Tucker, L.

2003 *Simplistic statistics*. Lincoln (UK) : Chalcombe Publications.

Tuggy, D.

1993 Ambiguity, polysemy, and vagueness. *Cognitive Linguistics* 4(3) : 273-290.

Valente, R.

2002 *La 'Lexicologie explicative et combinatoire' dans le traitement des unités lexicales spécialisées*. Thèse de doctorat. Université de Montréal.

Van Campenhoudt, M.

2000 De la lexicographie spécialisée à la terminographie : vers un 'métadictionnaire' ? In H. Béjoint & P. Thoiron (Eds.), *Le sens en terminologie* 127-152. Lyon : Presses universitaires de Lyon.

2001 Pour une approche sémantique du terme et de ses équivalents. *International Journal of Lexicography* 14(3) : 181-209.

2002a Lexicographie *vs* terminographie : quelques implications théoriques du projet DHYDRO. In H. Zinglé (Ed.), *Travaux du Lilla n° 4* 91-103. Université de Nice-Sophia Antipolis.

2002b Linguistique de corpus et étude des vocabulaires spécialisés. *Séminaire St-Denis, 8 janvier 2002, Université Paris 8* : *présentation non publiée* : http://www.termisti.refer.org/marcweb.htm

2005 Initier à la recherche de contextes d'attestation en langue spécialisée : une expérience didactique. In G. Williams (Ed.), *La linguistique de corpus* 297-306. Rennes : Presses universitaires de Rennes.

Vangehuchten, L.

2004 El uso de la estadística en la didáctica de las lenguas extranjeras con fines específicos : descripción del proceso de selección del léxico típico del discurso económico empresarial en español. *Actes de JADT 2004* : 1128-1135.

Vasilescu, F. & P. Langlais

2004 Désambiguïsation de corpus monolingues par des approches de type Lesk. *Actes de TALN 2004* : http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/taln2004-Fez/Vasilescu-Langlais.pdf

Venant, F.

2004 Polysémie et calcul du sens. *Actes de JADT 2004* : 1145-1156.

Veronis, J.

1998 A study of polysemy judgements and inter-annotator agreement. *Programme and advanced papers of the Senseval workshop, Herstmonceux Castle, England* : 2-4.

2001 Sense tagging : does it make sense ? *Paper presented at the Corpus Linguistics'2001 Conference, Lancaster, U.K.* http://www.up.univ-mrs.fr/veronis/pdf/2001-lancaster-sense.pdf

2003 Hyperlex : Cartographie lexicale pour la recherche d'informations. *Actes de TALN 2003* : 265-274.

2004a Quels dictionnaires pour l'étiquetage sémantique ? *Le Français Moderne* 2004(1) : 27-38.

2004b Hyperlex : lexical cartography for information retrieval. *Computer, Speech and Language* 18(3) : 223-252.

Victorri, B. & C. Fuchs

1992 Construire un espace sémantique pour représenter la polysémie d'un marqueur grammatical : l'exemple de *encore*. *Linguisticae Investigationes* XVI (1) : 125-153.

1996 *La polysémie*. *Construction dynamique du sens*. Paris : Hermès.

Victorri, B.

1997a La polysémie : un artefact de la linguistique ? *Revue de sémantique et pragmatique* 2 : 41-62.

1997b Modéliser les interactions entre une expression polysémique et son co-texte. In C. Guimier (Ed.), *Co-texte et calcul du sens* 233-245. Caen : Presses Universitaires de Caen.

Wandmacher, T.

2005 How semantic is Latent Semantic Analysis ? *Actes de RECITAL (TALN)* 2005 : 525-534.

Weber, M., R. Vos & H. Baayen

2000 Extracting the Lowest-Frequency Words : Pitfalls and Possibilities. *Computational Linguistics* 26(3) : 301-317.

Wehrens, R.

2004 *Introductory Statistics : reader*. Web tutorials in chemistry. http://www.webchem.science.ru.nl/Stat/stat.pdf

Welkenhuysen-Gybels, J. & G. Loosveldt

2002 *Regressieanalyse* : *een introductie in de multivariabelenanalyse*. Leuven : Acco.

Williams, G.

2002 In search of representativity in specialised corpora. Categorisation through collocation. *International Journal of Corpus Linguistics* 7(1) : 43-64.

Wüster, E.

1931 *Internationale Sprachnormung in der Technik* : *besonders in der Elektrotechnik*. Berlin : VDI-Verlag.

1968 *Dictionnaire multilingue de la machine-outil* : *notions fondamentales, définies et illustrées, présentées dans l'ordre systématique et l'ordre alphabétique* : *anglais-français*. London : Technical Press.

1991 *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. 3. Aufl. Bonn : Romanistischer Verlag.

Yarowsky, D.

1992 Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings COLING '92* : 454-460.

1994 Decision lists for lexical ambiguity resolution : application to accent restoration in Spanish and French. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* Las Cruces NM : 88-95.

1995 Unsupervised word sense disambiguation rivalling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Cambridge MA : 189-196.

2000 Hierarchical decision lists for Word Sense Disambiguation. *Computers and the Humanities* 30(1-2) : 179-186.

Zimina, M.

2004 Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. *Actes de JADT 2004* : 1195-1202.

- *Corpus*

Beauchet, J.

1996 *La rectification des pièces de révolution*. Collection Guides pratiques. Cluses : C.T.DEC

Kaufeld, M. & S. Torbaty

1999 *Rationalisation de l'usinage très grande vitesse*. Boulogne : Société Française d'Editions Techniques SOFETEC.

Sandvik Coromant

1997 *Techniques modernes d'usinage*. Sandviken (Suède) : AB Sandvik Coromant.

Schulz, H.

1997 *Fraisage à grande vitesse des matériaux métalliques et non métalliques*. Boulogne : Société Française d'Editions Techniques SOFETEC.

http://www.trametal.com

http://www.metal-industries.com

http://www.machine-outil.com

http://www.machine-outil.info

http://www.machpro.fr/magazine/default.htm

http://normach.wtcm.be/french/directives.html

http://ibn.be : EN 12417 (centres d'usinage) EN ISO 15641 (fraises pour usinage à grande vitesse) EN 12717 (perceuses) EN 12957 (machines d'électro-érosion) EN 13128 (fraises) EN 13218 (machines à meuler fixes)

- *Dictionnaires*

Kluwer

2001 Groot Polytechnisch Woordenboek – Grand Dictionnaire Polytechnique, Nederlands-Frans, français-néerlandais. Deventer/Anvers : Kluwer.

Nouveau Petit Robert

2001 Dictionnaire alphabétique et analogique de la langue française. Version électronique (CD-ROM) Version 2.0. Paris : Dictionnaires Le Robert/ VUEF.

http://www.sciences-en-ligne.com/Frames_Dictionary.asp

http://europa.eu.int/eurodicautom/login.jsp

http://membres.lycos.fr/baobab/techdico.html

http://www.granddictionnaire.com

http://www.m-w.com/home.htm

- *Logiciels*

MS Office

Textpad – Syn Text Editor – SciTE Text Editor

OmniPage Pro 11

Cordial 7 Analyseur : Synapse Développement, Toulouse

http://www.synapse-fr.com/

Abundantia Verborum : Speelman, D., Faculteit Letteren, Katholieke Universiteit Leuven

http://wwwling.arts.kuleuven.be/genling/abundant

http://wwwling.arts.kuleuven.be/av-tools/av-freq-doc.html

Lexico3 : outils de statistique textuelle : SYLED – CLA2T, Paris3

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

WordSmith : Scott, M., *Wordsmith Tools* version 3, Oxford : Oxford University Press.

http://www.lexically.net/wordsmith/

http://www.oup.com

Python 2.3.4. for Windows

http://www.python.org/

"R" Version 2.0.1. : The R Project for Statistical Computing

http://cran.r-project.org/

Summary

Polysemy in a technical lexicon
A quantitative study

This dissertation involves a semantic investigation into the domain of machining terminology in French. Building on a quantitative approach and corpus data (viz. a lemmatised corpus of French technical texts of about 1,7 million tokens), the investigation attempts to find out whether, and to what extent, pivotal lexical items are polysemous. Therefore, this study aims at developing a monosemy measure in order to quantify semantic analysis. Hence, the research question: is there a positive correlation between the typicality continuum of pivotal lexical items and the monosemy continuum ? Or conversely, is there any evidence for the hypothesis that the most typical or the most representative lexical items are not always the most monosemous items ?

- *Questioning the traditional dichotomy*

Linguistic tradition has it that communication in LSP (Language for Special Purposes) requires greater precision and univocity, and therefore aims at maximal “monosemy of terms”. The *first chapter* of this dissertation presents an overview of previous studies on specialised language and on semantic analysis, which are both characterised by a traditional dichotomy: “words versus terms” and “polysemy versus monosemy”. General language words can be polysemous, whereas specialised language terms are ideally monosemous. Recent studies, however, question the monosemy ideal of traditional terminology (Cabré 1991 and 1998 ; Temmerman 1997 and 2000a ; Gaudin 1993 and 2003). Analysis of specialised corpora (Eriksen 2002 ; Ferrari 2002) indeed confirms the polysemy of some lexical items, even inside a specialised domain. Furthermore, the traditional criteria for identifying and distinguishing monosemy, polysemy, homonymy and vagueness are not always reliable or convergent.

- *Alternative solution: continuum approach*

Since the traditional dichotomy is inappropriate for the analysis of specialised corpora, we decide to adopt a continuum approach. Typical lexical items are positioned on a typicality continuum and on a monosemy continuum, by means of their degree of typicality and their degree of monosemy. As a result of this continuum approach, the "monosemy thesis" of traditional terminology has to be paraphrased into a measurable and objective research question (*chapter 2*). If the traditional "monosemy thesis" holds, it will hold even more for the most typical lexical items of a technical lexicon. Hence, the main research question concerns the correlation between the typicality continuum and the monosemy continuum. The subsidiary research questions are concerned with the same correlation, but focus on separate wordclasses and separate subcorpora.

- *Double quantitative analysis: keywords and co-occurrences*

The implementation of the continuum approach, both in terms of degrees of typicality and of degrees of monosemy, requires a double quantitative analysis, explained in the methodological part of this dissertation. *Chapter 3* discusses the design of the technical corpus (1,7 million tokens) and the general language reference corpus of French journal articles (15,3 million tokens). *Chapter 4* presents two methodological approaches in order to establish the continuum of typicality ("Calculation of Specificities" and "Keywords Method"). To that end, the technical corpus is compared to the reference corpus. Although both approaches yield a similar list of keywords with a typicality coefficient, the Keywords Method is technically more efficient. The tool *Abundantia Verborum Frequency List Tool* implements the Keywords Method and generates a list of 4717 statistically significant "keywords", that is to say statistically significant typical lexical items of the technical corpus. Function words and proper nouns are filtered out. Typicality coefficients are then used to sort keywords by descending degree of typicality and to position them on the typicality continuum, ranging from most typical to least typical keywords.

Finally, *chapter 5* discusses the methodological aspects of the monosemy continuum. In order to calculate the degree of monosemy of the 4717 keywords, monosemy is implemented in terms of "semantic homogeneity". As a result, the degree of monosemy of a keyword is quantified in terms of the degree of formal overlap between the co-occurrents of the co-occurrents of the keyword (viz. the second order co-occurrents). If more second order co-occurrents are shared, the keyword is more homogeneous semantically and hence more monosemous. Monosemy degrees are then used to sort keywords by descending degree of monosemy (semantic homogeneity) and to position them on the monosemy continuum, ranging from most monosemous to least monosemous keywords. In

chapter 6, the monosemy measure is refined and yields a technical monosemy measure, weighted for the typicality of the second order co-occurrents.

- *Results and linguistic interpretations*

In *chapter 7*, the quantitative data are submitted to various statistical regression analyses. A simple linear regression analysis focuses on the impact of *typicality rank* (independent variable) on *monosemy rank* (dependent variable). A stepwise multiple regression analysis explores the impact of various independent variables on *monosemy rank*. The findings of the simple linear regression are confronted with the traditional "monosemy thesis". Contrary to what is expected, our technical corpus reveals a negative correlation between the typicality continuum and the monosemy continuum, as the most typical lexical items turn out to be less monosemous. The weighted technical monosemy measure yields similar results (negative correlation), although less convincing. Furthermore, the correlation between typicality rank and monosemy rank is not really linear and reveals a heteroskedasticity problem. The current technical solutions yield good results and confirm our research hypothesis (negative correlation). However, the results are not easy to interpret linguistically. Since the non-linear regression shows that less typical (or more general) items disturb the overall negative correlation, we decide to exclude those items. Indeed, the regression model does not account for these "general words", which are rather polysemous, despite their typicality rank. The 3210 remaining "technical" items do not show a heteroskedasticity problem and their results reveal a fairly good negative correlation between typicality and monosemy rank, with the most typical keywords being less monosemous. Yet, further research is required to point out whether, and to what extent, the "monosemy" of traditional terminology matches the degree of monosemy calculated by our monosemy measure and implemented in terms of semantic homogeneity.

The results of the stepwise multiple regression analysis confirm the findings of the simple regression analysis. The most significant factor is frequency rank in the technical corpus, with the most frequent items being most polysemous. The other significant factors are typicality rank, word length of the item (number of characters) and number of word classes the item belongs to.

Finally, *chapter 8* discusses the results of simple and multiple regression analyses on several subsets of the 4717 keywords (nouns, adjectives, verbs, adverbs) and on the keywords of the four subcorpora (electronic reviews, technical files, technical standards, handbooks). In particular the subcorpus of the technical standards is of great interest, since it contains mostly prescriptive and normative texts. The results of the main regression analyses (chapter 7) are confirmed both for the word classes and the subcorpora. For each word class, the findings reveal a negative correlation,

with the most typical items of a word class being less monosemous of that word class. The best correlation is found for the nouns and the worst correlation for the adverbs. Given the fact that nouns are usually well represented in a technical corpus, these findings confirm and corroborate our basic hypothesis. Furthermore, nouns, as opposed to adverbs, have stronger collocational and disambiguating mechanisms, which are clearly reflected in the monosemy measure, building on co-occurrence analysis. The negative correlation for some subcorpora is most convincing in the standards and in the handbooks. Most typical keywords in the standards turn out to be most polysemous in the standards and even in the entire technical corpus.

- *Perspectives and further research*

Building on a double large-scale quantitative analysis (viz. typicality continuum and monosemy continuum), this study provides quantitative and linguistic answers to semantic questions. However, new questions emerge during the research.

In this study, we focus on single word items, since it is rather difficult to determine the typicality degree of multiword expressions using the Keywords Method. Despite these technical limitations on the typicality level, multiword expressions should be submitted to a quantitative analysis on the semantic level, because they constitute a majority of the typical lexical items in a specialised corpus. Our quantitative semantic analysis can easily be conducted on multiword expressions, provided the monosemy measure builds on third order co-occurrences. Furthermore, most typical lexical items (e.g. *machine*), which turn out to be most heterogeneous semantically, typically show up in multiword expressions (e.g. *machine à usiner*), which will probably explain part of their semantic heterogeneity. Further research will explore the possible correlation between, on the one hand, polysemous and typical single word items and, on the other hand, the number of relevant multiword expressions.

We would like to refine our monosemy measure by incorporating more linguistic information, for example information on syntactic categories. It would also be interesting to complement our monosemy measure with cluster analyses, in order to group co-occurrents of a target word on the basis of the shared second order co-occurrents. These cluster analyses might yield more fine-grained semantic distinctions, complementary to the monosemy degree and monosemy rank.

Finally, we would like to test our double quantitative approach on other specialised corpora, in order to compare the results and to see if they also confirm our basic hypothesis. The quantitative semantic analysis could even be conducted on a general language corpus. We hope that our study and the double quantitative approach it encompasses, will elicit further research in quantitative semantics.

Samenvatting

De polysemie van technische woordenschat
Een kwantitatieve studie

Voorliggend proefschrift beoogt een semantisch onderzoek van de technische woordenschat in het domein van de metaalbewerkingsmachines. Aan de hand van kwantitatief corpusonderzoek (op basis van een gelemmatiseerd corpus Franse technische teksten van ongeveer 1,7 miljoen woorden), wordt bestudeerd of en in welke mate de typische woorden polyseem zijn. Deze studie heeft dan ook mede tot doel een monosemiemaat te ontwikkelen om de semantische analyse te kwantificeren. De centrale onderzoeksvraag gaat na of er een positieve correlatie bestaat tussen enerzijds het typiciteitscontinuum van typische woorden en anderzijds het monosemiecontinuum. Er wordt met name onderzocht of er enig kwantitatief-empirisch bewijs kan worden geleverd voor de hypothese dat de meest typische woorden of lexicale items niet altijd de meest monoseme zijn.

- *De traditionele dichotomie in vraag gesteld*

Vaktalige communicatie heeft vaak een grotere behoefte aan precisie en eenduidigheid, wat door de traditionele terminologie wordt opgevat als het ideaal van maximale monosemie. Het *eerste hoofdstuk* van dit proefschrift geeft een overzicht van voorgaand onderzoek met betrekking tot vaktaal en met betrekking tot semantische analyse. Beide onderzoeksdomeinen werden namelijk lange tijd gekenmerkt door een traditionele dichotomie: “woord versus term” en “polysemie versus monosemie”. Woorden uit de algemene taal kunnen meerdere betekenissen hebben, maar gespecialiseerde vaktermen zijn idealiter monoseem. Recente studies stellen deze traditionele monosemiestelling echter in vraag (Cabré 1991 en 1998 ; Temmerman 1997 en 2000a ; Gaudin 1993 en 2003). Uit analyse van vaktaal en van gespecialiseerde corpora (Eriksen 2002 ; Ferrari 2002) blijkt inderdaad dat sommige lexicale items polyseem zijn, zelfs binnen hetzelfde vakgebied. Bovendien zijn de

traditionele criteria voor het onderscheiden van monosemie, polysemie, homonymie en vaagheid niet altijd even betrouwbaar, noch leiden ze tot convergente resultaten.

- *Het continuum als alternatieve oplossing*

Het feit dat de traditionele dichotomie niet van toepassing blijkt te zijn bij de analyse van vaktaal noopt ons tot een alternatieve oplossing, die wordt opgevat als een dubbel continuum. De typische lexicale items van het technisch corpus worden in een typiciteitscontinuum en in een monosemiecontinuum gesitueerd, op basis van hun typiciteitsgraad en hun monosemiegraad. Om de traditionele monosemiestelling te verifiëren of te falsifiëren aan de hand van een continuum, dient deze stelling echter geherformuleerd te worden in een meetbare en objectieve onderzoeksvraag (*hoofdstuk 2*). Als de traditionele monosemiestelling wordt bevestigd in vaktalig corpusmateriaal, dan zal dit zeker gelden voor de meest typische of de meest specifieke lexicale items. De centrale onderzoeksvraag bestudeert bijgevolg de correlatie tussen het typiciteitscontinuum en het monosemiecontinuum in het technisch corpus. De bijkomende onderzoeksvragen onderzoeken deze correlatie voor typische lexicale items per woordsoort en per subcorpus.

- *Een dubbele kwantitatieve analyse: "keywords" en co-occurrenties*

De implementatie van het typiciteits- en het monosemiecontinuum, aan de hand van typiciteits- en monosemiegraden, veronderstelt een dubbele kwantitatieve analyse, die wordt behandeld in het methodologisch deel van dit proefschrift. *Hoofdstuk 3* beschrijft de samenstelling van het technisch corpus (1,7 miljoen woorden) en van het referentiecorpus algemene taal bestaande uit Franse journalistieke teksten (15,3 miljoen woorden).

In *hoofdstuk 4* worden de twee methodes besproken om het typiciteitscontinuum op te bouwen, namelijk de "berekening van specifieke items" (*calcul des spécificités*) en de "sleutelwoorden-" of "keywordsmethode" (*Keywords Method*). Hiertoe wordt het technisch corpus vergeleken met het referentiecorpus. Beide methodes leveren gelijkaardige resultaten op, namelijk een lijst met "keywords" en hun typiciteitscoëfficiënt. Onze voorkeur gaat evenwel uit naar de keywordsmethode, omdat die technisch efficiënter is. Het softwarepakket Abundantia Verborum Frequency List Tool implementeert de keywordsmethode en genereert een lijst met 4717 statistisch significante keywords of lexicale items die typisch zijn voor het technisch corpus. Functiewoorden en eigennamen zijn hier wel uitgefilterd. Aan de hand van de typiciteitscoëfficiënten worden de keywords dan gesorteerd in dalende volgorde van specificiteit of typiciteit. Zo kunnen de keywords makkelijk gesitueerd worden in het typiciteitscontinuum, gaande van de meest typische tot de minst typische keywords.

Hoofdstuk 5 behandelt de methodologische aspecten van het monosemiecontinuum. Om de monosemiegraad van de 4717 keywords te berekenen, wordt “monosemie” beschouwd als “semantische homogeniteit”. Op die manier kan er een cijfer geplakt worden op monosemie en kan de monosemiegraad van een typisch lexicaal item gekwantificeerd worden als de mate van formele overlap van de co-occurrenten van zijn co-occurrenten (i.e. de co-occurrenten van de tweede orde). Als er meer co-occurrenten van de tweede orde gemeenschappelijk zijn en dus worden gedeeld door co-occurrenten van de eerste orde, dan is het basiswoord semantisch meer homogeen en bijgevolg meer monoseem. De berekende monosemiegraden worden dan gebruikt om de keywords te sorteren en te situeren in een monosemiecontinuum, gaande van de meest monoseme (of semantisch homogene) tot de minst monoseme keywords. In *hoofdstuk 6* wordt de monosemiemaat verder verfijnd en wordt er ook een technische monosemiemaat uitgewerkt, die een weging voorziet in functie van de typiciteit (of de techniciteit) van de co-occurrenten van de tweede orde.

- *Resultaten en linguïstische interpretaties*

De kwantitatieve gegevens, zowel voor typiciteit als voor monosemie, worden vervolgens in *hoofdstuk 7* onderworpen aan een aantal statistische regressieanalyses. Een enkelvoudige lineaire regressieanalyse bestudeert de impact van typiciteitsrang (onafhankelijke variabele) op monosemierang (afhankelijke variabele). Een stapsgewijze meervoudige regressieanalyse onderzoekt de gecombineerde impact van meerdere onafhankelijke variabelen op monosemierang. De resultaten van de enkelvoudige lineaire regressieanalyse worden dan geconfronteerd met de traditionele monosemiestelling. In tegenstelling tot wat men zou verwachten, is er in het technisch corpus een negatieve correlatie tussen het typiciteitscontinuum en het monosemiecontinuum: de meest typische lexicale items blijken de minst monoseme te zijn. De gewogen technische monosemiemaat leidt tot gelijkaardige resultaten (negatieve correlatie), hoewel iets minder overtuigend. Bovendien blijkt ook dat de correlatie tussen typiciteitsrang en monosemierang niet helemaal lineair is en dat er dus een probleem van heteroscedasticiteit is.

De gebruikelijke technische oplossingen voor heteroscedasticiteit bieden goede resultaten. Ze bevestigen onze onderzoekshypothese (negatieve correlatie), maar linguïstisch zijn deze resultaten vrij moeilijk te interpreteren. Aangezien de niet-lineaire regressie aantoont dat de minst typische (of de meest algemene) items een storend effect hebben op de globale negatieve correlatie, worden deze items uitgesloten uit de analyse. Het regressiemodel blijkt inderdaad niet goed te werken voor deze “algemene woorden”, die sowieso vrij polyseem (of vaag) zijn, ongeacht hun typiciteitsrang. Voor de 3210 overblijvende “technische” typische lexicale items is er geen heteroscedasticiteitsprobleem meer en is er bovendien een vrij goede negatieve correlatie tussen typiciteits- en monosemierang, waarbij de meest typische

items de minst monoseme zijn. Verder onderzoek is echter vereist om na te gaan of en in welke mate de “monosemie” van de traditionele terminologie overeenkomt met de monosemiegraad die wordt berekend door onze monosemiemaat en die wordt beschouwd als semantische homogeniteit.

De resultaten van de stapsgewijze meervoudige regressieanalyse bevestigen de bevindingen van de enkelvoudige regressieanalyse. De meest significante factor is de frequentierang in het technisch corpus: de meest frequente typische items blijken de meest polyseme te zijn. De andere significante factoren zijn typiciteitsrang, woordlengte van het item (in aantal letters) en aantal woordsoorten waartoe het item behoort.

Hoofdstuk 8 bespreekt de resultaten van de enkelvoudige en de meervoudige regressieanalyses voor bepaalde subsets van de 4717 keywords (substantieven, adjectieven, werkwoorden en bijwoorden) en voor de keywords van de subcorpora (tijdschriften, technische fiches, normen, handboeken). Vooral het subcorpus van de normen is bijzonder interessant, omdat het grotendeels prescriptieve en normatieve teksten bevat. De bevindingen van de basisanalyses (hoofdstuk 7) worden inderdaad bevestigd, zowel voor de verschillende woordsoorten als voor de subcorpora. Voor elke woordsoort blijkt er een negatieve correlatie te zijn, waarbij de meest typische lexicale items per woordsoort de minst monoseme zijn van die woordsoort. De substantieven vertonen de beste correlatie, de bijwoorden de slechtste. Rekening houdend met het feit dat substantieven meestal erg goed vertegenwoordigd zijn in een technisch corpus, bevestigen en versterken deze resultaten onze basishypothese. In tegenstelling tot bijwoorden, hebben substantieven bovendien sterke collocatie- en desambigueringsmechanismen. Deze komen duidelijk tot uiting in de monosemiemaat, die gebaseerd is op co-occurrentieanalyse. Bij de subcorpora is de negatieve correlatie het sterkst in de normen en in de handboeken. De meest typische woorden in de normen blijken de meest polyseme, zowel in de normen zelf als in het volledige technisch corpus.

- *Perspectieven voor verder onderzoek*

Aan de hand van een dubbele en grootschalige kwantitatieve analyse (i.e. typiciteits- en monosemiecontinuum), geeft deze dissertatie kwantitatieve en linguïstische antwoorden op een belangrijke semantische vraag. Er duiken evenwel ook nieuwe vragen op tijdens het onderzoek.

In deze studie concentreren wij ons op enkelvoudige woorden. Het is immers vrij moeilijk om de typiciteitsgraad te bepalen van meerwoordige lexicale eenheden aan de hand van de keywordsmethode. Ondanks de technische beperkingen met betrekking tot typiciteit, is een kwantitatieve analyse op semantisch vlak (monosemiegraad) wel aangewezen voor deze meerwoordige lexicale eenheden. Ze

vertegenwoordigen immers het grootste deel van de typische lexicale items van een technisch corpus. De kwantitatieve semantische analyse die in dit proefschrift is uitgewerkt kan perfect worden uitgevoerd op meerwoordige lexicale eenheden, indien de monosemiemaat wordt gebaseerd op co-occurrenten van de derde orde. Trouwens, de meest typische lexicale items, zoals bijvoorbeeld *machine*, die semantisch het meest heterogeen blijken te zijn, komen zeer vaak voor in meerwoordige lexicale eenheden (*machine à usiner*). Dit verklaart waarschijnlijk gedeeltelijk hun semantisch heterogeen karakter. Verder onderzoek zal ondermeer nagaan of er een correlatie bestaat tussen enerzijds polyseme typische enkelvoudige items en anderzijds het aantal relevante meerwoordige lexicale eenheden waarvan ze deel uitmaken.

We denken verder ook aan een verfijning van onze monosemiemaat door meer linguïstische informatie op te nemen, zoals bijvoorbeeld woordsoortinformatie. Het zou bovendien interessant zijn om onze monosemiemaat aan te vullen met clusteranalyses, om op die manier de co-occurrenten van een basiswoord te clusteren of te groeperen op basis van hun gemeenschappelijke co-occurrenten (i.e. gedeelde co-occurrenten van de tweede orde). Deze clusteranalyses zouden eventueel kunnen leiden tot preciezere en fijnmazigere semantische onderscheidingen, complementair met betrekking tot de monosemiegraad en –rang.

Tenslotte zouden we onze dubbele kwantitatieve en scalaire aanpak graag testen op andere gespecialiseerde corpora, om de resultaten te vergelijken en na te gaan of ze onze basishypothese bevestigen. De kwantitatieve semantische analyse kan trouwens ook op een corpus algemene taal worden uitgevoerd. Wij hopen in elk geval dat ons doctoraatsonderzoek, met zijn dubbele kwantitatieve methode, kan leiden tot verder onderzoek in het domein van de kwantitatieve semantiek.

Glossaire linguistique

Bruit	cooccurrents qui ne sont pas pertinents
Cooccurrence	présence simultanée des occurrences de deux mots différents dans un contexte donné ou dans une fenêtre d'observation donnée
Cooccurrent	mot qui « cooccurre » avec le mot analysé (mot de base) ou qui apparaît dans son voisinage, c'est-à-dire dans la même séquence, dans le même paragraphe, etc. ; les cooccurrents sont généralement caractérisés par leur présence simultanée dans une fenêtre d'observation (*span*), par exemple de 5 mots à gauche et à droite
Fenêtre d'observation	distance autour du mot analysé, exprimée en nombre de mots à gauche et à droite, par exemple une fenêtre d'observation de 5 mots à gauche et de 5 mots à droite du mot analysé (= *span*)
Hétérogénéité sémantique	une unité lexicale est hétérogène sémantiquement si ses cooccurrents de premier ordre (ou ses contextes d'usage) sont différents entre eux et s'ils appartiennent à des champs sémantiques nettement différents
Homogénéité sémantique	une unité lexicale est homogène sémantiquement si ses cooccurrents de premier ordre (ou ses contextes d'usage) sont similaires entre eux et s'ils appartiennent au même champ sémantique
Mot	une suite de caractères entre deux espaces et/ou signes de ponctuation ; plus particulièrement, une unité lexicale (ou grammaticale) de la langue générale

Occurrence	chaque apparition d'une unité linguistique dans un texte ou dans un corpus, sur le plan de la parole ou du discours (= *token*) ; le nombre total d'occurrences d'un corpus indique l'étendue du corpus et correspond au nombre total d'unités linguistiques dénombrées (formes graphiques ou lemmes), dont la plupart sont récurrentes
Recoupement	les occurrences d'une liste se recoupent de manière significative ou présentent un recoupement important si plusieurs d'entre elles figurent plusieurs fois dans la liste ; les cooccurrents de deuxième ordre (ou cc) d'un mot de base se recoupent beaucoup, s'ils figurent plusieurs fois dans la liste des cc, donc s'ils sont partagés par plusieurs cooccurrents de premier ordre de ce mot de base
Sens	signifié linguistique d'un signe ou d'une unité linguistique au niveau de la parole ou du discours (à l'opposé de la signification au niveau de la langue)
Signe	unité linguistique constituée d'un signifiant matériel, concret et observable (la forme graphique ou sonore), et d'un signifié abstrait (le contenu sémantique)
Signifiant	forme concrète et observable d'un signe (la forme graphique ou sonore)
Signification	signifié linguistique d'un signe ou d'une unité linguistique en ce qui concerne le système de la langue (à l'opposé du sens en ce qui concerne la parole)
Signifié	contenu sémantique d'un signe
Silence	cooccurrents pertinents pour l'analyse, mais que l'on n'a pas relevés
Spécificité	une unité lexicale (ou unité grammaticale) est spécifique quand elle est représentative d'une section d'un corpus par rapport au corpus entier ou représentative d'un corpus de langue spécialisée par rapport à un corpus de référence de langue générale (= mot-clé)

Terme	unité lexicale (simple ou complexe) de la langue spécialisée, représentant un concept à l'intérieur d'un domaine spécialisé, par exemple *usinage* ou *machine à usiner* (= unité terminologique, unité spécialisée)
Type	chaque unité linguistique différente, considérée sur le plan de la langue (= *type*) ; le nombre de types dans un corpus correspond au nombre d'unités linguistiques différentes (formes graphiques ou lemmes)
Type-Token Ratio	rapport entre le nombre d'unités linguistiques différentes (types) d'un corpus et le nombre total d'unités linguistiques (occurrences), qui permet de mesurer la richesse lexicale du corpus ou la diversité de son vocabulaire (= TTR)
Unité grammaticale	unité linguistique simple ou complexe, variable ou invariable, dont la fonction syntaxique est plus importante que sa fonction sémantique ; un déterminant, un pronom, une préposition, une conjonction, un adverbe (autre qu'un adverbe en *–ment*) ou un verbe auxiliaire (= mot fonctionnel, mot « vide »)
Unité lexicale	unité linguistique simple ou complexe, généralement variable, dont la fonction sémantique est plus importante que sa fonction syntaxique : un nom, un adjectif qualificatif, un adverbe en -ment ou un verbe (sauf auxiliaire) (= mot « plein »)
Unité polylexicale	une unité lexicale complexe, constituée de plusieurs unités lexicales simples, par exemple *machine à usiner*

Glossaire statistique

Analyse de régression (I)	une analyse de régression simple permet d'étudier l'impact d'une variable indépendante sur une variable dépendante ; par exemple, on peut étudier l'impact du rang de spécificité sur le rang de monosémie, mais aussi l'impact du niveau de formation sur la chance (ou la probabilité) de trouver un emploi
Analyse de régression (II)	une analyse de régression multiple permet d'étudier l'impact combiné de plusieurs variables indépendantes sur une seule variable dépendante ; par exemple, on peut étudier l'impact du rang de spécificité et du rang de fréquence sur le rang de monosémie, mais aussi l'impact du niveau de formation et du nombre d'années d'expérience sur la chance (ou la probabilité) de trouver un emploi
Coefficient de corrélation	le coefficient de corrélation sert à quantifier la relation entre deux variables et il est exprimé par un nombre décimal entre 1 et -1 ; plus il s'approche de 1, plus la corrélation positive est forte ; plus il s'approche de -1, plus la corrélation négative est forte ; un coefficient de corrélation de 0 ou s'approchant de 0 indique une absence de corrélation entre les deux variables étudiées
Droite de régression	la droite de régression visualise le résultat d'une analyse de régression simple ; elle traverse le nuage des points en minimisant la distance entre chaque point et la droite (= droite des moindres carrés)
Hypothèse nulle	si l'hypothèse nulle est vérifiée, les phénomènes observés sont dus au hasard et ne sont pas statistiquement significatifs

Rapport de vraisemblance	le rapport de vraisemblance (ou *log-likelihood ratio*) est une mesure statistique pour déterminer statistiquement la pertinence d'une différence ou d'une relation, observée dans un corpus par exemple (= LLR)
Résidus	les résidus d'une analyse de régression sont les erreurs d'estimation commises lorsqu'on prédit les valeurs de la variable dépendante à partir des valeurs de la variable indépendante ; sur la visualisation (*plot*), le résidu d'une observation correspond à la distance verticale entre le point noir qui visualise l'observation et la droite de régression
Seuil de significativité	le seuil de significativité, en fonction d'une valeur p, indique le seuil auquel un phénomène observé est statistiquement significatif ou pertinent ; le seuil de significativité le plus courant est associé à une valeur p inférieure à 0,05 ($p < 0,05$), ce qui veut dire qu'il y a 5% de chances que le phénomène observé soit dû au hasard ; à partir de ce seuil de significativité (p.ex. 0,05), on rejette l'hypothèse nulle d'indépendance (= seuil de rejet)
Valeur p	la valeur p ou la valeur de probabilité indique la significativité statistique d'un modèle de régression ou d'une variable indépendante, sous l'hypothèse nulle d'indépendance ; la valeur p la plus courante est inférieure à 0,05 ($p < 0,05$)
Variable dépendante	variable dont on souhaite expliquer ou prédire la variation dans une analyse de régression (= variable à expliquer, variable expliquée, variable prédite) ; par exemple, le rang de monosémie
Variable indépendante	variable qui permet d'expliquer ou de prédire la variation d'une variable dépendante dans une analyse de régression (= variable explicative, variable prédictive) ; par exemple, le rang de spécificité, le rang de fréquence, la classe lexicale, etc.

Variation expliquée	le pourcentage de variation expliquée est le résultat d'une analyse de régression simple ou multiple ; il représente le pourcentage de variation de la variable dépendante (ou expliquée) que l'on pourra expliquer ou prédire à partir de la variation de la variable indépendante (ou explicative) ; le pourcentage de variation expliquée est aussi appelé le coefficient de détermination (= R^2) ; il correspond au carré du coefficient de corrélation

---

Gqtest	test statistique de Goldfeld-Quandt qui permet de constater l'hétéroscédasticité ou l'homoscédasticité des résidus d'une analyse de régression en fonction d'une valeur p (une valeur $p < 0{,}05$ indique l'hétéroscédasticité)
LLR	*Log-Likelihood Ratio* (= le rapport de vraisemblance)
R^2	le pourcentage de variation expliquée (ou le coefficient de détermination)
Valeur p	la valeur de probabilité
VD	la variable dépendante
VI	la variable indépendante